



UNIVERSITY OF
BIRMINGHAM

META-ANALYSIS OF RISK PREDICTION STUDIES

by

IKHLAAQ AHMED BSc. MSc.

A thesis submitted to the

University of Birmingham

for the degree of

DOCTOR OF PHILOSOPHY

Public Health, Epidemiology & Biostatistics

School of Health and Population Sciences

College of Medical and Dental Sciences

University of Birmingham

January 2015

UNIVERSITY OF
BIRMINGHAM

University of Birmingham Research Archive

e-theses repository

This unpublished thesis/dissertation is copyright of the author and/or third parties. The intellectual property rights of the author or third parties in respect of this work are as defined by The Copyright Designs and Patents Act 1988 or as modified by any successor legislation.

Any use made of information contained in this thesis/dissertation must be in accordance with that legislation and must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the permission of the copyright holder.

META-ANALYSIS OF RISK PREDICTION STUDIES

Ikhlaaq Ahmed BSc. MSc.

Abstract

A statistical model that quantifies the relationship between outcome risk and one or more predictors (prognostic factors) is regarded as a risk prediction model. In this thesis I identify and demonstrate the methodological challenges of meta-analysing risk prediction models using either aggregate data or individual patient data (IPD).

Firstly, a systematic review of published breast cancer models is performed, to summarise their content and performance using published aggregate data. It is found that models were not available for comparison, as they were not presented in most cases. This reveals poor reporting standards in primary studies. To address this issue, meta-analysis of individual patient data (IPD) is proposed and a systematic review performed to examine articles that develop and/or validate a risk prediction model using IPD from multiple studies. This identifies that most articles only use the IPD for model development, and thus ignore external validation, and also ignore clustering of patients within studies. In response to these issues, IPD is obtained from one of the articles and used as a case study. This article uses parathyroid hormone (PTH) assay (a continuous variable) to predict postoperative hypocalcaemia after thyroidectomy. It is shown statistically that ignoring clustering is inappropriate, as it ignores potential between-study heterogeneity in discrimination and calibration performance, potentially producing misleading inferences for clinical practice. In particular, post-model probabilities are shown to calibrate better if heterogeneity is accounted for by tailoring model results to the prevalence in the intended population.

This dataset was also used to evaluate an imputation method for dealing with missing thresholds when IPD are unavailable, and the simulation results indicate the approach performs well, though further research in other datasets is also required.

This thesis therefore makes a positive contribution towards meta-analysis of risk prediction models to improve clinical practice. Nonetheless, many methodological issues are outlined for further research.

ACKNOWLEDGEMENTS

This thesis was made possible by funding from the MRC Midlands hub for trials and methodology research and the department of public health, epidemiology and biostatistics.

Firstly, I would like to sincerely thank Richard Riley for providing me with the opportunity to undertake this research and for his supervision, encouragement, guidance and unwavering support during this thesis. I am specifically grateful for how he has helped develop my career path and improve my research skills, and for the large amount of time he has spent reading various drafts of the thesis in the past few years and providing constructive feedback. I would also like to thank Jon Deeks and Lucinda Billingham who have similarly guided and advised me throughout this thesis and have given up their valuable time to listen and encourage. I would like to thank everyone I have worked with during my thesis who has provided encouragement and support.

My love and appreciation goes to my family, who have helped me through difficult times and encouraged me when necessary. In particular, I would like to thank my Mum, Dad, Owais and all the rest of my family members! Yes, this is why I have been so busy in the past few years! I would also like to thank all my friends, whose encouragement has been invaluable at times.

Finally, I would like to thank God, for helping me overcome every hurdle in life and this thesis and for giving me strength through every difficult situation. I credit all the good parts of the thesis to you; it would not have been possible without your blessings!

Table of Contents

CHAPTER 1: INTRODUCTION.....	1
1.1 Overview.....	1
1.2 Prognosis research	2
1.2.1 Why is prognosis research important?	3
1.2.2 What is meant by a predictive factor and a risk prediction model?	3
1.3 Basic statistical concepts for risk prediction model.....	5
1.3.1 Model development.....	5
1.3.1.1 Logistic regression.....	5
1.3.1.2 Cox model	6
1.3.2 Model validation	8
1.3.2.1 Performance statistics	11
1.3.2.2 Discrimination	11
1.3.2.3 Calibration	12
1.4 What is a systematic review and meta-analysis?	14
1.4.1 Systematic review	14
1.4.2 Meta-analysis	16
1.4.2.1 Why is meta-analysis important?	16
1.4.2.2 Aggregate data meta-analysis.....	17
1.4.2.3 Fixed-effect model.....	18
1.4.2.4 Random-effects model.....	19
1.4.2.5 Random-effects meta-analysis and the prediction interval.....	21
1.4.3 Individual patient data (IPD) meta-analysis.....	22
1.4.3.1 Why is it potentially useful?	23
1.4.3.2 Limitations of IPD	24
1.5 Aims and thesis summary	25
1.5.1 Chapter 2	25
1.5.2 Chapter 3	26
1.5.3 Chapter 4	26
1.5.4 Chapter 5	26
1.5.5 Chapter 6	27
1.5.6 Chapter 7	27

CHAPTER 2: RISK PREDICTION OF BREAST CANCER: SYSTEMATIC REVIEW AND META-ANALYSIS OF AGGREGATE DATA FROM PUBLISHED STUDIES

.....	28
2.1 Introduction.....	28
2.2 Breast cancer.....	29
2.3 Objectives of the systematic review	30
2.4 Methods	31
2.4.1 Searching and screening.....	31
2.4.1.1 Searching for potentially relevant articles	31
2.4.1.2 Inclusion and Exclusion Criteria	31
2.4.1.3 Screening	32
2.4.2 Data extraction	33
2.4.2.1 Quality of included studies	33
2.4.2.2 Performance (validation) statistics	34
2.4.2.3 Model parameter estimates.....	36
2.4.3 Meta-analysis	36
2.4.3.1 C statistic and E/O statistic.....	36
2.4.3.2 Model parameter estimates.....	37
2.5 Results.....	37
2.5.1 Search results.....	37
2.5.2 Study quality and reporting.....	40
2.5.3 Articles developing a prediction model	40
2.5.4 Articles externally validating a model	47
2.5.4.1 Qualitative summary	47
2.5.4.2 Meta-analysis of external validation statistics.....	51
2.5.4.3 Meta-analysis of the refitted parameter estimates	55
2.6 Discussion.....	60
2.6.1 Limitations and strengths of the review	61
2.6.2 The best model for predicting breast cancer risk?.....	62
2.6.3 Reporting standards.....	63
2.6.4 Next steps for prediction model research in breast cancer.....	64
2.6.5 Difficulties of doing an aggregate data meta-analysis of risk prediction models	66
2.6.6 What would be the advantages of IPD?	66
2.6.7 Next steps for the thesis	67
2.6.8 What this chapter adds?.....	68

CHAPTER 3: SYSTEMATIC REVIEW OF ARTICLES DEVELOPING OR VALIDATING A PREDICTION MODEL USING IPD FROM MULTIPLE STUDIES

.....	69
3.1 Introduction.....	69
3.2 Methods	70
3.2.1 Identifying potentially relevant articles.....	70
3.2.2 Inclusion and exclusion criteria.....	71
3.2.3 Data extraction and in-depth evaluation of articles.....	72
3.3 Results.....	75
3.3.1 Classification results	75
3.3.2 Background information	76
3.3.3 Research objectives	76
3.3.4 Identifying studies	79
3.3.5 Obtaining IPD from multiple studies	79
3.3.6 Details of IPD obtained	81
3.3.7 Missing data	82
3.3.8 Model development.....	83
3.3.8.1 Analysis method	83
3.3.8.2 Heterogeneity of predictor effects	85
3.3.8.3 Handling of continuous predictors	85
3.3.8.4 Need for standardisation.....	86
3.3.8.5 Strategy for inclusion of predictors	86
3.3.8.6 Reporting of developed model	87
3.3.9 Model validation	90
3.3.9.1 Internal, external or internal-external validation	90
3.3.9.2 Reporting of validation performance criteria	92
3.3.10 Dealing with those studies not willing / able to provide their IPD	96
3.3.11 Conclusions and limitations	96
3.4 Discussion.....	98
3.4.1 Limitations of the review	99
3.4.2 Methodological challenges to consider	100
3.4.3 Areas for improvement - Recommendations	102
3.4.3.1 Recommendation One: Allow for different baseline risks in each of the IPD studies.....	103
3.4.3.2 Recommendation Two: Implement a framework that uses internal- external cross-validation.....	105
3.4.4 Conclusions	107

3.4.5 Findings in context for the thesis	108
3.4.6 What this chapter adds?.....	108
CHAPTER 4: UNSTRATIFIED VERSUS META-ANALYTIC APPROACHES FOR EVALUATING PREDICTIVE TEST ACCURACY USING IPD: PART 1 - DISCRIMINATION	109
4.1 Introduction and objectives.....	109
4.2 Summary of Noordzij article	110
4.3 Summary of the IPD available	113
4.4 Methods	114
4.4.1 Noordzij analysis: unstratified approach.....	115
4.4.2 Meta-analysis approaches	117
4.4.2.1 Univariate meta-analysis for sensitivity and specificity.....	118
4.4.2.2 Bivariate meta-analysis for sensitivity and specificity	119
4.4.2.3 C statistic	120
4.5 Results.....	121
4.5.1 Comparison of mean values	122
4.5.2 Sensitivity and Specificity.....	124
4.5.2.1 Comparison of discrimination results at 0-20 minutes.....	124
4.5.2.2 Comparison of sensitivity and specificity at 1-2 hours and 6 hours.....	127
4.5.3 C statistic	131
4.6 Discussion.....	137
4.6.1 Limitations	139
4.6.2 What this chapter adds?.....	139
CHAPTER 5: UNSTRATIFIED VERSUS META-ANALYSIS APPROACHES FOR EVALUATING PREDICTIVE TEST ACCURACY USING IPD: PART II – CALIBRATION AND CHOICES OF THRESHOLD	141
5.1 Introduction and objectives.....	141
5.2 Statistical methods for summarising calibration.....	143
5.2.1 Approaches for assessing calibration	143
5.2.1.1 Approach 1: Unstratified	144
5.2.1.2 Approach 2: Univariate summary sensitivity and specificity, and univariate summary prevalence	145
5.2.1.3 Approach 3: Univariate summary sensitivity and specificity, lower predicted prevalence	147
5.2.1.4 Approach 4: Univariate summary sensitivity and specificity, upper predicted prevalence	147
5.2.1.5 Approach 5: Univariate summary sensitivity and specificity, study- specific prevalence	148

5.2.1.6 Approach 6: Internal-external validation.....	148
5.3 Results I - Calibration	150
5.3.1 Comparison of PPV and NPV estimates	151
5.3.2 Comparison of the calibration of predicted values.....	156
5.3.2.1 0-20 minutes	156
5.3.2.2 1-2 hours	163
5.3.3 Meta-analysis of E/O estimates	166
5.3.3.1 0-20 minutes time-point.....	166
5.3.3.2 1-2 hours' time-point.....	169
5.3.3.3 Meta-analysis results for Approaches 1, 2 and 5 for the thresholds 60-80	170
5.3.4 Approach 6, 'internal-external' cross-validation.....	172
5.4 Results II -What is the best threshold?	177
5.4.1 Illustration of how meta-analysis can change the best threshold decision based on sensitivity and specificity alone	177
5.4.2 What is the best threshold?.....	178
5.4.2.1 Harm and benefit I.....	179
5.4.2.2 Harm and benefit II	181
5.4.2.3 Harm and benefit III	183
5.5 Discussion	185
5.5.1 Key findings	186
5.5.2 What this Chapter adds?.....	189
CHAPTER 6: A SIMULATION STUDY TO EMPIRICALLY EVALUATE AN IMPUTATION METHOD FOR DEALING WITH MISSING THRESHOLDS IN META-ANALYSIS OF A PREDICTIVE TEST	190
6.1 Introduction.....	190
6.2 Methodology of the imputation approach.....	192
6.2.1 Details of the method	192
6.2.2 Example of how this method works for the McLeod study [144]	196
6.2.3 Example of the imputation approach	198
6.3 Methods for a simulation study of the imputation method	200
6.3.1 Dataset.....	201
6.3.2 Simulation of 1000 datasets with missing threshold.....	202
6.3.3 Meta-analysis of missing data and imputed data	204
6.3.4 Programming code	205
6.3.5 Statistical measures	206
6.4 Results of the simulation study of the imputation method	206

6.4.1 Scenario I: Probability of missing equals 0.5 for all thresholds.....	206
6.4.1.1 Pooled estimates	207
6.4.1.2 Standard error of pooled estimates	207
6.4.1.3 Between-study variances	207
6.4.2 Scenario II: 60% and 70% thresholds have probability missing of 0.1, and the remaining thresholds have probability missing of 0.5	213
6.4.2.1 Pooled estimates	213
6.4.2.2 Standard error of pooled estimates	213
6.4.2.3 Between-study variances	213
6.4.3 Scenario III: Data not missing at random, thresholds with sensitivity<0.90 having a probability of 0.5 of being missing, but otherwise are reported.....	219
6.4.3.1 Pooled estimates	219
6.4.3.2 Standard error of pooled estimates	219
6.4.3.3 Between-study variances	220
6.4.4 Scenario IV: Missing not at random, with the first and last thresholds always being present. The data has a probability of missing equals 0.5 for the other thresholds if the specificity is <0.80	226
6.5 Discussion	229
6.5.1 Strengths and limitations of simulation study	229
6.5.2 Strengths and limitations of the imputation method	230
6.5.3 Multivariate meta-analysis model approach	231
6.5.4 Other methods	232
6.5.5 Conclusion.....	233
6.5.6 What further research is needed?	234
6.5.7 What this chapter adds?.....	234
CHAPTER 7: DISCUSSION	235
7.1 Overview of the thesis	235
7.2 Key findings and recommendations	237
7.3 Main messages for the risk prediction field	239
7.4 Further research	241
7.5 Conclusions.....	244
APPENDIX A.....	245
APPENDIX B.....	250
APPENDIX C.....	257
APPENDIX D.....	270
List of References.....	274

List of Illustrations

Figure 2.1: PRISMA diagram for systematic review	38
Figure 2.2: Meta-analysis of the C statistic for breast cancer risk prediction models....	52
Figure 2.3: Meta-analysis of the E/O ratio for breast cancer risk prediction models.....	53
Figure 2.4: Meta-analysis of age at menarche for Gail 2 models.....	56
Figure 2.5: Meta-analysis of one biopsy \times age categorised interaction term (age <50) for Gail 2 models.....	57
Figure 2.6: Meta-analysis of one biopsy \times age categorised interaction term (age ≥ 50) for Gail 2 models.....	57
Figure 2.7: Meta-analysis of one affected first degree relatives \times age at first live birth \times age <20 interaction term for Gail 2 models.....	58
Figure 2.8: Meta-analysis of one affected first degree relatives \times age at first live birth \times age 20-24 interaction term for Gail 2 models.....	58
Figure 2.9: Meta-analysis of one affected first degree relatives \times age at first live birth \times age 25-29 interaction term for Gail 2 models.....	59
Figure 2.10: Meta-analysis of one affected first degree relatives \times age at first live birth \times age ≥ 30 interaction term for Gail 2 models.....	59
Figure 3.1: Graph showing distribution of the year of publication of the 15 articles identified.....	75
Figure 3.2: IPD obtained and total IPD studies desired in the seven articles using a literature review to identify relevant studies	80
Figure 3.3: Horn et al. ROC curve.....	93
Figure 3.4: Number of articles that have cited the Royston et al. paper, by each year	107
Figure 4.1: ROC curves comparing the unstratified and meta-analysis approaches for 0-20 minutes	126
Figure 4.2: ROC curves comparing the unstratified and meta-analysis approaches for 1-2 hours	129
Figure 4.3: ROC curves comparing the unstratified and meta-analysis approaches for 6 hours	130
Figure 4.4: 0-20 minutes unstratified ROC curve	132
Figure 4.5: 1-2 hours unstratified ROC curve	132
Figure 4.6: 6 hours unstratified ROC curve	133
Figure 4.7: 0-20 minutes study-specific ROC curves.....	134
Figure 4.8: 1-2 hours study-specific ROC curves	135
Figure 4.9: 6 hours study-specific ROC curves.....	135
Figure 5.1: Meta-analysis of E/O for the 3 different approaches at the 0-20 minute's time-point	169
Figure 5.2: Meta-analysis of E/O for the 3 different approaches at the 1-2 hours' time-point.....	170
Figure 5.3: Meta-analysis of E/O values from 'internal-external' cross validation approach for threshold 65% and time-point 1-2 hours.....	177
Figure 6.1: Illustrating the Imputation method using the McLeod study.....	197
Figure 6.2: ROC curve comparing before and after imputation data	200

Figure 6.3: ROC curves for all five studies for the 1-2 hour time-point from Noordzij data	201
Figure 6.4: Box plots of Sensitivity, Logit-Sensitivity, S.E's for Logit-Sensitivity and Tau-squared for Sensitivity against Threshold %, comparing Non-Imputed and Imputed data (Scenario I).....	210
Figure 6.5: Box plots of Specificity, Logit-Specificity, S.E's for Logit-Specificity and Tau-squared for Specificity against Threshold %, comparing Non-Imputed and Imputed data (Scenario I).....	211
Figure 6.6: S.E's of Logit-Sensitivity for each threshold (Scenario I) in each of the 1000 datasets, before and after imputation.....	212
Figure 6.7: S.E's of Logit-Specificity for each threshold (Scenario I) in each of the 1000 datasets, before and after imputation.....	212
Figure 6.8: Box plots of Sensitivity, Logit-Sensitivity, S.E's for Logit-Sensitivity and Tau-squared for Sensitivity against Threshold %, comparing Non-Imputed and Imputed data (Scenario II)	216
Figure 6.9: Box plots of Specificity, Logit-Specificity, S.E's for Logit-Specificity and Tau-squared for Specificity against Threshold %, comparing Non-Imputed and Imputed data (Scenario II)	217
Figure 6.10: S.E's of Logit-Sensitivity for each threshold (Scenario II) in each of the 1000 datasets, before and after imputation.....	218
Figure 6.11: S.E's of Logit-Specificity for each threshold (Scenario II) in each of the 1000 datasets, before and after imputation.....	218
Figure 6.12: Box plots of Sensitivity, Logit-Sensitivity, S.E's for Logit-Sensitivity and Tau-squared for Sensitivity against Threshold %, comparing Non-Imputed and Imputed data (Scenario III).....	223
Figure 6.13: Box plots of Specificity, Logit-Specificity, S.E's for Logit-Specificity and Tau-squared for Specificity against Threshold %, comparing Non-Imputed and Imputed data (Scenario III).....	224
Figure 6.14: S.E's of Logit-Sensitivity for each threshold (Scenario III) in each of the 1000 datasets, before and after imputation.....	225
Figure 6.15: S.E's of Logit-Specificity for each threshold (Scenario III) in each of the 1000 datasets, before and after imputation.....	225

List of Tables

Table 2.1 Models and their outcomes.....	43
Table 2.2: Risk prediction models - reporting and analysis characteristics	44
Table 2.3: Breast cancer risk prediction models - comparison of predictors included ..	46
Table 2.4: Breast cancer risk prediction models - validation statistics	49
Table 2.5: What chapter two adds	68
Table 3.1: Background information for each article.....	77
Table 3.2: Summarising the IPD	81
Table 3.3: Details of missing data in the IPD projects and how it was handled	83
Table 3.4: Model development.....	88
Table 3.5: Model validation methods and results reported in the articles	94
Table 3.6: Limitations and methodological problems noted in discussion of the 15 articles.....	97
Table 3.7: Methodological challenges, extending those identified for prognostic factors by Abo-Zaid et al.	101
Table 3.8: Areas for improvement and recommendations	102
Table 3.9: What chapter three adds?	108
Table 4.1: Summary of the 6 IPD studies.....	113
Table 4.2: Two by two table for each threshold	116
Table 4.3: Summary of the PTH values before and after thyroidectomy pooled across the 6 studies and in each study separately	121
Table 4.4: Sensitivity and Specificity of PTH assay in predicting postoperative hypocalcaemia for 0-20 minutes: unstratified, univariate and bivariate results	125
Table 4.5: Table for Youden's statistic comparing the different approaches for 0-20 minutes	127
Table 4.6: Sensitivity and Specificity of PTH assay in predicting postoperative hypocalcaemia for 1-2 hours: unstratified, univariate and bivariate results.....	128
Table 4.7: Table for Youden's statistic comparing the univariate approach to the unstratified approach for 1-2 hours	129
Table 4.8: Sensitivity and Specificity of PTH assay in predicting postoperative hypocalcaemia for 6 hours: unstratified, univariate and bivariate results	130
Table 4.9: Table for Youden's statistic comparing the two approaches for 6 hours....	131
Table 4.10: C statistic for each study, within each time-point	134
Table 4.11: C statistic random-effects meta-analysis	136
Table 4.12: C statistic meta-analysis across all studies	136
Table 4.13: What Chapter four adds?	140
Table 5.1: Summary of the 6 IPD studies.....	151
Table 5.2: PPV and NPV for 0-20 minutes	152
Table 5.3: PPV and NPV for 1-2 hours	154
Table 5.4: PPV and NPV for 6 hours	155
Table 5.5: E/O ratios for all the patients combined using the unstratified and univariate meta-analysis approaches at the 0-20 minute's time-point	157
Table 5.6: E/O ratios using 5 approaches for McLeod study at the 0-20 minute's time-point.....	159

Table 5.7: E/O ratios using 5 approaches for Lo study at the 0-20 minute's time-point	160
Table 5.8: E/O ratios using 5 approaches for Lombardi study at the 0-20 minute's time-point.....	162
Table 5.9: E/O ratios for all the patients combined using the unstratified and univariate meta-analysis approaches at the 1-2 hours' time-point	163
Table 5.10: E/O ratios using 5 approaches for McLeod study at the 1-2 hours' time-point.....	164
Table 5.11: E/O ratios using 5 approaches for Lam study at the 1-2 hours' time-point	165
Table 5.12: Meta-analysis values for E/O for various thresholds at the 0-20 minutes and 1-2 hour time-points	172
Table 5.13: Sensitivity and specificity values when 'internal-external' cross-validation approach compared to the univariate approach for the 1-2 hours' time-point	174
Table 5.14: E/O ratios for all combinations of excluded studies.....	175
Table 5.15: Sensitivity and Specificity of PTH assay in predicting postoperative hypocalcaemia for 1-2 hours using the same two studies that provided data for the 6 hours analysis; univariate meta-analysis results.....	178
Table 5.16: Table for Youden's statistic comparing the two approaches for 1-2 hours, using the two studies that have provided data for the 6 hour time point.....	178
Table 5.17: Largest NPV for 0-20 minutes	180
Table 5.18: 1-2 hours.....	181
Table 5.19: 6 hours	181
Table 5.20: Worse to send someone home incorrectly for 0-20 minutes.....	182
Table 5.21: Worse to send someone home incorrectly for 1-2 hours.....	182
Table 5.22: Worse to send someone home incorrectly for 6 hours	182
Table 5.23: Harm and benefit equal and worse to send someone home for 0-20 minutes	183
Table 5.24: Harm and benefit equal and worse to send someone home for 1-2 hours	184
Table 5.25: Harm and benefit equal and worse to send someone home for 6 hours....	184
Table 5.26: What chapter five adds?	189
Table 6.1: 2 by 2 table for thresholds 60%-70%, showing what the values are after imputation.....	197
Table 6.2: Summarising the studies evaluating the Apgar score.....	198
Table 6.3: Univariate meta-analysis results for data before and after imputation.....	199
Table 6.4: Sensitivity and Specificity of PTH assay in predicting postoperative hypocalcaemia for 1-2 hours: univariate and bivariate results	202
Table 6.5: Scenario I, Probability of missing equals 0.5 for all thresholds.....	208
Table 6.6: Scenario II, 60% and 70% thresholds have probability missing of 0.1, and the remaining thresholds have probability missing of 0.5.....	214
Table 6.7: Scenario III; Data not missing at random, thresholds with sensitivity<0.90 having a probability of 0.5 of being missing, but otherwise are reported	221
Table 6.8: Scenario IV; missing not at random, with the first and last thresholds always being present. The data has a probability of missing equals 0.5 for the other thresholds if the specificity is <0.80.....	227
Table 6.9: What chapter six adds?.....	234
Table 7.1: Key findings and recommendations	238
Table 7.2: Further research recommendations	244

List of Equations

(Equation 1.1)	6
(Equation 1.2)	6
(Equation 1.3)	7
(Equation 1.4)	7
(Equation 1.5)	8
(Equation 1.6)	12
(Equation 1.7)	12
(Equation 1.8)	13
(Equation 1.9)	13
(Equation 1.10)	18
(Equation 1.11)	18
(Equation 1.12)	18
(Equation 1.13)	19
(Equation 1.14)	19
(Equation 1.15)	20
(Equation 1.16)	20
(Equation 1.17)	20
(Equation 1.18)	20
(Equation 1.19)	20
(Equation 1.20)	21
(Equation 1.21)	21
(Equation 1.22)	21
(Equation 1.23)	21
(Equation 1.24)	21
(Equation 1.25)	22
(Equation 2.1)	35
(Equation 2.2)	35
(Equation 4.1)	115
(Equation 4.2)	116
(Equation 4.3)	116
(Equation 4.4)	117
(Equation 4.5)	118
(Equation 4.6)	118
(Equation 4.7)	119
(Equation 4.8)	120
(Equation 4.9)	121
(Equation 5.1)	144
(Equation 5.2)	144
(Equation 5.3)	145
(Equation 5.4)	145
(Equation 5.5)	145
(Equation 5.6)	145
(Equation 5.7)	146

(Equation 5.8)	146
(Equation 5.9)	179
(Equation 6.1)	192
(Equation 6.2)	192
(Equation 6.3)	193
(Equation 6.4)	193
(Equation 6.5)	193
(Equation 6.6)	193
(Equation 6.7)	194
(Equation 6.8)	194
(Equation 6.9)	194
(Equation 6.10)	194
(Equation 6.11)	195
(Equation 6.12)	195
(Equation 6.13)	196

List of Abbreviations

IPD	Individual Patient Data
IMPACT	International Mission for Prognosis and Analysis of Clinical Trials in Traumatic Brain Injury
QOL	Quality Of Life
OS	Overall Survival
ROC	Receiver Operating Characteristic
PPV	Positive Predictive Value
NPV	Negative Predictive Value
E/O	Expected/Observed
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta- Analyses
RR	Relative Risk
OR	Odds Ratio
HR	Hazard Ratio
BMI	Body Mass Index

CHAPTER 1: INTRODUCTION

1.1 Overview

A statistical model that quantifies the relationship between outcome risk and one or more predictive factors is regarded as a ***risk prediction model***; such models allow the risk of an outcome to be predicted for a new patient based on their specific predictive values [1-4]. For example, a risk prediction model has been developed to predict the risk of an unfavourable six month outcome [5] for patients with traumatic brain injury by using the predictive factors of age, motor score, pupillary reactivity, CT characteristics, and laboratory parameters. Risk prediction models have many potential uses. In particular, they may help determine treatment for a patient and help clinicians make decisions based on the patient's predicted risk. For example, those patients at high-risk may be given therapies which are expensive or have harmful potential side-effects.

Although many risk prediction models are proposed in the medical literature, only a small proportion seems to be used in practice [6, 7]. In particular, after their development, risk prediction models require validation in a different dataset (external to that used for model development) but many models do not perform well when they are externally validated and most are never externally validated.

Meta-analysis of individual patient data (IPD) offers an innovative opportunity for risk prediction model research. In this situation, patient-level data are available from multiple studies for the purposes of model development and validation, which provides

unique opportunities. In particular, models can be developed using data from a subset of studies and assessed on data from the remaining studies. For example, the IMPACT consortium developed a prognostic model for mortality and unfavourable outcome in traumatic brain injury by using IPD shared by 11 studies (8509 patients), with successful external validation using IPD from another large study (6681 patients) [5]. This allowed the developed model to be validated successfully and increased the chance it would be used in practice.

The overall aims of the thesis are to evaluate, apply and develop methods for risk prediction research within the context of meta-analysis, with particular consideration of the benefits of IPD over and above other meta-analysis approaches (e.g. by using published model results). The full aims and overview of the thesis are provided at the end of this chapter, but to begin with some key aspects of prognosis and prediction that are fundamental to the thesis are introduced.

1.2 Prognosis research

Prognosis means to predict or estimate the probability of a future outcome. In medical research, prognosis is related to the risk or probability of an individual developing an outcome over a specified time period (based on their profile/characteristics, which are termed prognostic or predictive factors). The outcome can vary from death to disease progression in those with disease, or onset of a disease if the patient is healthy to begin with, along with changes in quality of life (QOL) or pain.

1.2.1 Why is prognosis research important?

Prognosis research is important in a medical setting due to its ability to guide and inform individuals about the estimated future course of their illness, or to inform healthy individuals of their risk of developing an illness or disease. It allows the investigation of relationships between future outcomes and baseline health to help improve health [8]. This makes a huge difference in clinical practice as it allows doctors to predict, and thus understand, the likely course of an illness in a particular individual based on the individuals' characteristics. For example, a breast cancer screening tool can be used to determine if an individual has a probability of developing the cancer. This can help guide decisions regarding treatments (what changes the individual can make, or which treatment is the better one to use, or if a change in treatment is required etc.).

There are quite a few similarities between aetiological and prognostic research, with the design often being a cohort study where patients have factors measured at baseline and their outcomes recorded prospectively over time. However aetiological research is often more focused on causality, whereas prognosis research is more focused on prediction. Predicting an outcome is different to explaining the cause of an outcome, as every causal factor is a predictor, but not every predictor is a cause [9].

1.2.2 What is meant by a predictive factor and a risk prediction model?

Predictive factors are ultimately any patient characteristic or measure that can be used to predict the most likely clinical outcome for that patient [10]. These can include simple factors like age, sex, stage of disease, or more complex factors such as unusual genetic

mutations. These can be used to predict the response to treatment, the most appropriate treatment or to aid patient counselling.

Due to the clinical importance of prognostic factors there is an ongoing need for individual primary studies to identify suitable prognostic factors that can be used in practice. The next stage is to combine prognostic factors together in a statistical model, to identify individuals who have a high risk of developing an adverse outcome, so they can be subject to early preventative strategies and possible treatment. For example, an individual that appears healthy but is found to have a high risk of developing cardiovascular disease could be recommended to modify their lifestyle choice (e.g. smoking, exercise etc.) or be prioritised for a clinical investigation which could lead to early diagnosis of an underlying condition (e.g. diabetes, high blood pressure).

Therefore, there is a growing interest in risk prediction modelling for the purpose of prognostic risk assessments [11-13], where a statistical model is used to estimate the risk of a future outcome based on one or more characteristics. A risk prediction model is also referred to as a prognostic model (when the outcome risk is for patients with a defined disease) or more generally a clinical prediction model (used for both diseased and non-diseased settings). Similarly the word 'model' is often replaced with 'score', 'tool', 'index' or 'rule'. Patient characteristics are simple predictors, but can be termed as prognostic factors, risk factors, prognostic markers and prognostic variables.

A risk prediction model uses predictors (covariates) to estimate the absolute risk of an outcome using an individual's predictive profile. A diagnostic prediction model predicts the absolute risk that an outcome is present (i.e. disease) and a prognostic prediction

model predicts the absolute risk that an outcome will occur in a specified time period (i.e. relapse in the next year) [11].

1.3 Basic statistical concepts for risk prediction model

There are two main phases for risk prediction model research: model development (including internal validation using the same data or data source) and external validation (using new data from a different data source) [12, 14, 15].

1.3.1 Model development

A risk prediction model is developed to enable the estimation of the outcome risk using either a single or multiple predictive factors [11]. Commonly used statistical models for risk prediction are the Cox regression and logistic regression models [16]. There are two strategies often utilised in selecting predictors for a model. First is a selection procedure, which is usually forwards or backwards (and eliminates those not statistically significant from the model), and the second is simply including every predictor in the model [11]. Logistic and Cox regression are now described.

1.3.1.1 Logistic regression

Logistic regression is used for prediction of the probability of occurrence of an event by a particular time, by fitting data to a logistic probability function. This, like many other forms of regression analysis, uses several predictor variables that are either continuous or categorised. For example, the probability that a woman has first onset of breast cancer could be predicted by her age, Body Mass Index (BMI), smoking habits, alcohol consumption and family history of breast cancer.

We begin by explaining the logistic function, from which logistic regression comes from. If we call the probability of an event p , then we start by taking the odds, $p / (1-p)$ and then take the log of the odds to get $\ln (p / (1-p))$. Now this logistic function is useful as it maps p from a value between 0 and 1 to $\pm \infty$.

$$\ln \left(\frac{p}{1-p} \right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k, \text{ where } 0 \leq k \leq \infty$$

(Equation 1.1)

Here, β_0 is called the intercept term, and β_1, β_2 etc are called the regression coefficients of x_1, x_2 etc. The intercept term is the value of $\ln \left(\frac{p}{1-p} \right)$ when all other variables are equal to zero, also known as the baseline risk. Each regression coefficient describes the size of the risk increase (the risk is quantified on the logit / log odds scale) for a 1-unit increase in the predictor. A positive coefficient means that an increase in the variable increases the probability of the outcome whereas a negative coefficient means that a decrease in the variable decreases the probability of the outcome. The β s relate to the increase in log odds, or equivalently log odds ratios. Once the model has been estimated, it can be rearranged to give a predicted probability for a new patient by:

$$\hat{p} = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k)}$$

(Equation 1.2)

1.3.1.2 Cox model

Cox regression model is used to analyse survival data, where the exact time of the event is important. It can describe the associations of different variables with survival; adjust for other confounding factors to look at the relationship of a particular variable and

survival; and predict the prognosis of an individual based on their characteristics (variables in the model, i.e. age, sex etc.).

With a single prognostic factor, the Cox proportional hazards model can be written as:

$$h_1(t) = h_0(t)\exp(\beta X)$$

(Equation 1.3)

Where $X=1$ is a binary prognostic factor, $\exp(\beta)$ here gives the hazard ratio, $h_0(t)$ is the baseline hazard and its distribution is difficult to specify. A simple choice would be to use the exponential distribution, so $h_0(t) = \lambda$ for $t \geq 0$; or a much more flexible approach would be the Weibull distribution, with $h_0(t) = k\lambda(\lambda t)^{k-1}$ for $t \geq 0$.

If the primary interest is in the comparison of groups (i.e. hazard ratio), rather than estimating the actual hazard rate in each group, Cox [17] proposed a semi-parametric approach that estimates the hazard ratio but makes no assumption about the baseline hazard.

Given several covariates of interest, the Cox model can be written as:

$$h_1(t) = h_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \dots)$$

(Equation 1.4)

The quantity $\beta_1 X_1 + \beta_2 X_2 + \dots$ is sometimes known as the risk score or the prognostic index, and categorising this creates risk groups (e.g. low, medium or high-risk etc.).

Risk groups can be shown by plotting the categories of the prognostic index on Kaplan Meier curves. The International Prognostic Index (IPI) is used to determine low risk

(91% OS), intermediate risk (78% OS) and high risk (53% OS) [18] for patients with aggressive non-Hodgkin's lymphoma.

If the baseline hazard is known, equation (1.4) can be used to make survival predictions over time by the following:

$$S(t) = S_0(t)^{\exp(\beta_1 X_1 + \beta_2 X_2 + \dots)}$$

(Equation 1.5)

Where $S_0(t)$ is the baseline survival function that corresponds to the baseline hazard.

1.3.2 Model validation

Model validation is important as it shows whether a risk prediction model is accurate. It is not regarded as being sufficient to show it works well in the development dataset alone [11], as usually the model performs well in the development data, but this may not carry over to a different population. This can be due to differing patient characteristics, model over fitting or non-inclusion of important predictors [11].

The purpose of validation is to show that the model is accurate in the population of individuals for whom it is intended. Predictive performance of a model is often optimistic when assessed using the development data (internal validation), so it is important to validate the model using individuals that were not in the development data and in a different setting (external validation) [19], i.e. using a dataset from a different study which is assessing the same outcome.

To validate the performance of a risk prediction model, it is important to assess calibration and discrimination. Calibration compares observed and expected (predicted)

event rates for groups of patients. Discrimination is the ability of the risk prediction model to distinguish between those patients who do or do not get the event. An example to illustrate the difference between calibration and discrimination is a model that predicts every Premier League football team has a three out of twenty chance of being relegated: as 3 teams are always relegated, this model will have perfect calibration overall. However, this model does not discriminate between teams, i.e. the predicted probability of survival is always 3 out of 20, as it does not take in to account any prognostic factors such as their league results from previous years, their home and away form, or whether this is their first year in the premier league. So the discrimination will be poor in this instance.

Model performance should ideally be assessed using an external dataset; this assesses generalizability, i.e. transporting the model to assess performance when taken to different populations [14]. There are three validation approaches discussed [14]:

1. Internal validation approach often involves splitting the data (i.e. by studies or number of patients) in to two parts (ratios such as 2:1 or 4:1) [14]. The risk prediction model is developed on the first proportion and then validated on the second proportion; due to the datasets being similar this can give over optimistic results. Internal validation only provides information on how well the model performs in the same population as it was developed on. Techniques such as cross-validation or bootstrapping can also be used toward internal validation.

Bootstrapping

Bootstrapping reflects the procedure of sampling from an approximating distribution [13]. A bootstrap sample can be taken from the underlying distribution which reflects the original data and are usually the same size; approximately 200 bootstrap samples can be enough to obtain steady estimates [13]. In the context of risk prediction model validation, the model is developed in each bootstrap sample and then validated in that same sample and the original data; the difference between the two validations is regarded as optimism, this optimism is then taken away from the performance of the original model in the original dataset [13].

2. Internal-external cross-validation (IECV) involves developing the risk prediction model on a proportion of a dataset (regarded as the ‘internal’ data) and then validating it on the remaining proportion (regarded as the ‘external’ data); this process is then repeated. For example, we can split the data in to five proportions, with each one fifth proportion being left out in turn as the validating dataset, thus allowing all the patients to take part in both the development and the validation of the risk prediction model [13]. The overall performance is usually regarded as the average of all the validations (across all omitted datasets).

3. Temporal validation, this is similar to internal validation, but the difference is that the data is split by time, thus being regarded as external in time [14]. This is independent of the development stage and is a prospective evaluation of the model; it is regarded as being in-between internal and external validation.

4. External validation examines the generalizability of the model and must use new data from a similar or different population [14]. The new dataset must contain the same prognostic factors recorded that the risk prediction model contains; otherwise the developed model cannot be implemented.

1.3.2.1 Performance statistics

The validation must determine the model's ability to differentiate between patients with different outcomes (discrimination) and show the agreement among observed and predicted risks in groups of individuals with similar risk predictions (calibration) [11].

1.3.2.2 Discrimination

The concordance (C) statistic is most often used to measure performance of a prediction model to indicate its discriminative ability [20]. It is a rank order statistic for predictions against the true outcomes and does not take in to account the errors in calibration in terms of the differences between average outcomes [13]. For binary outcomes the area under the Receiver Operating Characteristic (ROC) curve is identical to the C statistic, which plots sensitivity against 1-specificity [20]. C statistic also exists for survival outcomes [21].

Sensitivity and Specificity

If the risk prediction model equation is dichotomised (i.e. the prognostic index is split into two groups), or if a single predictor ('test') is evaluated, then sensitivity and specificity can be derived, which are measures of discrimination. Sensitivity (Equation 1.6) relates to a predictor's ability to identify a patient with the outcome correctly, i.e. the predictor result will be positive for a patient with the outcome [22]. Specificity

(Equation 1.7) relates to a predictor's ability to identify a non-diseased patient correctly, i.e. a predictor result will be negative for a patient without the outcome [22]. Both are either reported as proportions or percentages.

$$\text{Sensitivity} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}}$$

(Equation 1.6)

$$\text{Specificity} = \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}}$$

(Equation 1.7)

1.3.2.3 Calibration

When more than two predicted probabilities are possible (i.e. when the single predictor or prognostic index is left on a continuous scale), a graphical assessment of calibration is possible by plotting the predicted outcome risk on the x-axis and the observed outcome risk on the y-axis, and a perfect calibration would result in prediction being on the 45 degrees line [20]. This calibration plot can be understood by an intercept a which indicates whether the predictions are too high or too low (this is known as calibration-in-the-large which should be zero) as well as a calibration slope b , which should be equal to 1 [20]. When the model is developed $a=0$ and $b=1$ for regression models, but at validation calibration-in-the-large problems can occur and b can be smaller than 1 indicating the model is over estimating the true risk [20]. Calibration plots for survival can be done by plotting observed risks in Kaplan-Meier plots, against expected risks from the derived $S(t)$ function. A calibration plot does not purely assess calibration (in the same way that E/O assesses calibration only); despite the name it also shows some

aspects of discrimination by visually displaying the range of predicted risks across the range of 0 to 1.

Predictive values

The positive predictive value (PPV) (Equation 1.8) of a predictor is the probability that a patient with a positive result does truly have the outcome [22]. The negative predictive value (NPV) (Equation 1.9) is the probability that a patient with a negative result does truly not have the outcome [22]. Both predictive values are usually reported as proportions or percentages.

$$PPV = \frac{(\text{sensitivity}) * (\text{prevalence})}{(\text{sensitivity}) * (\text{prevalence}) + (1 - \text{specificity}) * (1 - \text{prevalence})}$$

(Equation 1.8)

$$NPV = \frac{(\text{specificity}) * (1 - \text{prevalence})}{(\text{specificity}) * (1 - \text{prevalence}) + (1 - \text{sensitivity}) * (\text{prevalence})}$$

(Equation 1.9)

E/O statistic

Expected (E) events can be calculated using PPV and NPV for a predictive test and the observed (O) events are those that are known. By dividing the expected by the observed we can get a statistic that tells us how well this model predicts and calibrates. An E/O value of over 1 indicates that the model is over predicting, a value of less than 1 indicates the model is under predicting and a value of 1 implies that there is perfect prediction and calibration.

1.4 What is a systematic review and meta-analysis?

In this thesis, the development and validation of risk prediction models will be considered in the context of systematic reviews and meta-analysis. Systematic reviews are increasingly important in the process of establishing the performance and clinical effectiveness of a risk prediction model. Many risk prediction models are developed, and systematic reviews and their subsequent meta-analysis allow the opportunity to gain a better overview and understanding of how well a risk prediction model actually performs by synthesising results from different studies.

1.4.1 Systematic review

A systematic review is a transparent framework for identifying, appraising, summarising and (if appropriate) synthesising research evidence from multiple studies of the same risk prediction model. Systematic reviews can examine quantitative or qualitative evidence. It should aim to be based on a protocol that can easily be replicated [23]. A systematic review of high quality should always aim to identify all the published and unpublished evidence that is related to the research question. There should be pre-defined inclusion criteria for the studies that will be systematically reviewed, with there being a method to assess the quality of those reports and studies that have been included. In an unbiased manner, the results or findings from those studies and reports should be synthesised. The review should aim to interpret the results or findings and present them in an impartial summary [23].

Systematic reviews are needed in risk prediction research because there may be many studies published for a single risk prediction model, or many different models specifically proposed for that disease and outcome area (i.e. breast cancer and survival).

Many of these studies may give unclear and confusing results and a lack of clarity in the findings makes it difficult to comprehend for clinicians and healthcare decision makers. Each study may offer little insight to the research question, but it is hoped that when a systematic review is performed, we are able to gain a clearer picture of the key findings or results [23]. A systematic review is important as otherwise non-systematic reviews are performed, e.g. literature/narrative/critical review/commentary. The problem with these types of reviews is that they do not have a protocol that they set and follow, so their findings are usually very difficult to replicate. A larger problem with these types of reviews is that small study effects are not being picked up, as different conclusions were reached by researchers using the same research base [23].

The systematic review process typically involves defining a particular question or aim, then a literature review with a transparent search strategy. Potential biases that may occur in this search are publication bias [23-25], language bias and selection bias. Searching for unpublished material, also known as the grey literature is therefore very important, as it may help reduce the effect of publication bias [23]. Whilst searching the literature it is common to contact the key authors of a study to ask for clarification on the article if required. Once the literature has been searched, the studies will be assessed based on inclusion criteria and for those that have been included, the full text papers will be obtained and usually an assessment of study quality is made. A data extraction form is then used to extract the findings or results from all the studies, the findings or results are then combined and this is known as evidence synthesis. When quantitative data is extracted about the risk prediction model (e.g. model predictor estimates or

validation performance statistics) then a meta-analysis might be performed [23], which gives a statistical summary of the predictors or model performance.

1.4.2 Meta-analysis

The formal definition of meta-analysis is ‘the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings’ [26].

This has now become a corner stone of evidence based medicine due to its ability to combine the results of individual studies. Meta-analysis, for example, could show whether a risk prediction model has a good or poor model performance by pooling validation statistics across studies [27].

A well conducted systematic review is essential for a good meta-analysis [28]. There are checklists available for the assessment of the quality of reporting of systematic reviews (note: quality of reporting is not the same as quality of conduct), with the PRISMA [29, 30] statement (quality of reporting of meta-analyses) being particularly recommended, though this is specific to intervention research rather than prediction modelling.

1.4.2.1 Why is meta-analysis important?

Meta-analysis is used to synthesise study results (predictor estimates or validation statistics), with explicit criteria for study inclusion or exclusion from the meta-analysis [31]. Meta-analysis helps reduce problems of interpretation due to the variability of observed effect estimates across trials (i.e. due to sampling variation, caused by the limited sample sizes of individual studies) [31]. It is generally used to quantify effect sizes and their uncertainty which is represented by a mean value and its 95% confidence

interval. One of the main purposes of meta-analysis is to facilitate synthesis of large numbers of study results.

A good conduct criterion for a meta-analysis would involve the meta-analysis being specified in a formal protocol [31]. There would be a compilation of complete set of studies and common validation statistics would be identified. There would be a standardised data extraction protocol, with an analysis allowing for sources of variation [31]. A sensitivity analysis to assess study quality or impact of each study on the results would usually be performed.

There are different approaches and methods that exist for meta-analysis; such as the vote counting method, combining p-values method, but the method used the most is combining estimates of effect using either a fixed-effect model or a random-effects model [31]. Different summary measures can be used depending on the measure of interest; for example, the odds ratio, relative risk, or hazard ratios for predictor effects, and c statistics, sensitivity or specificity, and calibration slope for model performance. Regardless, similar approach is taken where the estimate from each study is weighted by the precision of the estimate [28], as now described.

1.4.2.2 Aggregate data meta-analysis

An aggregate data meta-analysis essentially involves taking the value of the main effect estimate (validation or model predictor estimates) in a study along with its standard error and using it to perform a meta-analysis.

An aggregate data meta-analysis typically uses two models: the fixed-effect model or the random-effects model.

1.4.2.3 Fixed-effect model

A fixed-effect model assumes that there is no between-study heterogeneity and that all the studies are estimating the exact same predictor or validation performance; this may not be reasonable as studies often differ in design and population. By assuming a single (fixed) predictor/performance effect across studies, the pooled estimate then gives the best estimate of this single predictor/performance effect. The fixed-effect model can be written as follows, where Y_i is the predictor/validation effect estimate for each study ($i = 1 \dots n$). We interpret the coefficient β to represent the best estimate of a common effect across studies, and s_i^2 denotes $Var(Y_i)$ and is assumed known:

$$Y_i = \beta + e_i$$

$$e_i \sim N(0, s_i^2)$$

(Equation 1.10)

For ratios such as Relative Risk (RR), Odds Ratio (OR) or Hazard Ratio (HR); $Y_i = \ln(RR)$, $\ln(OR)$, or $\ln(HR)$. To calculate the weighted mean, each study needs to have a weight assigned to it:

$$W_i = \frac{1}{s_i^2}$$

(Equation 1.11)

s_i^2 is the within-study variance for study i . The weighted mean is calculated:

$$M = \frac{\sum_{i=1}^k W_i Y_i}{\sum_{i=1}^k W_i}$$

(Equation 1.12)

It is the effect size multiplied by the weight divided by the sum of the weights. The variance of this weighted mean is:

$$V_M = \frac{1}{\sum_{i=1}^k W_i}$$

(Equation 1.13)

The estimated standard error is the square root of the variance:

$$SE_M = \sqrt{V_M}$$

(Equation 1.14)

These results are equivalent to maximum likelihood solutions.

1.4.2.4 Random-effects model

The random-effects model allows the between study variability in effect to be accounted for. Therefore, it is now estimating a different effect for each study. This approach assumes a distribution of effects across studies and each study can then have a different effect. The pooled estimate then gives the estimate of the average effect across the studies. The random effect point estimates are usually similar to the fixed-effect one. However, the 95% confidence interval are usually wider [31] than the fixed-effect model, with studies being given a more equal weighting in a random-effects synthesis than a fixed-effect model. Equation (1.10) can be extended to a random-effects model (1.15), where Y_i is the effect estimate for each study ($i = 1 \dots k$). Now, β is the average effect from the distribution of effects across studies, and τ^2 the between-study variance, where $\theta_i = \beta + u_i$ and s_i^2 is still assumed known:

$$Y_i = \beta + u_i + e_i$$

$$u_i \sim N(0, \tau^2)$$

$$e_i \sim N(0, s_i^2)$$

(Equation 1.15)

τ^2 is calculated by:

$$\tau^2 = \frac{Q - df}{C}$$

(Equation 1.16)

Where

$$Q = \sum_{i=1}^k W_i Y_i^2 - \frac{(\sum_{i=1}^k W_i Y_i)^2}{\sum_{i=1}^k W_i}$$

(Equation 1.17)

$$df = k - 1$$

(Equation 1.18)

Here k is the number of studies, and:

$$\sum_{i=1}^k W_i - \frac{\sum_{i=1}^k W_i^2}{\sum_{i=1}^k W_i}$$

(Equation 1.19)

The same notations as fixed-effect will be used but with an asterisk to indicate it is the random-effects model. The weight for each study is:

$$W_i^* = \frac{1}{V_{Y_i}^*}$$

(Equation 1.20)

$V_{Y_i}^*$ is the within-study variance for study i plus τ^2 :

$$V_{Y_i}^* = V_{Y_i} + \tau^2$$

(Equation 1.21)

The weighted mean is calculated:

$$M^* = \frac{\sum_{i=1}^k W_i^* Y_i}{\sum_{i=1}^k W_i^*}$$

(Equation 1.22)

It is the effect size multiplied by the weight divided by the sum of the weights. The variance of this weighted mean is:

$$V_{M^*} = \frac{1}{\sum_{i=1}^k W_i^*}$$

(Equation 1.23)

The estimated standard error is the square root of the variance:

$$SE_{M^*} = \sqrt{V_{M^*}}$$

(Equation 1.24)

1.4.2.5 Random-effects meta-analysis and the prediction interval

Following a random-effects meta-analysis, researchers often use the average effect and its confidence interval. To consider the potential effect when applied to a new study is

also important, as this could differ from the average, and this can be assessed by calculating a prediction interval. The prediction interval gives a range of values for the predicted effect in a new study [32].

The 95% prediction interval is approximately:

$$\hat{\mu} - t_{k-2}\sqrt{\hat{\tau}^2 + SE(\hat{\mu})^2}, \hat{\mu} + t_{k-2}\sqrt{\hat{\tau}^2 + SE(\hat{\mu})^2}$$

(Equation 1.25)

Where $\hat{\mu}$ is the estimate of the average estimate (i.e. of the predictor/performance statistic) across studies from the random-effects meta-analysis, $SE(\hat{\mu})$ is the standard error of $\hat{\mu}$, $\hat{\tau}$ is the estimate of between-study standard deviation, t_{k-2} is the $100(1 - \alpha/2)$ percentile of the t-distribution with $k-2$ degrees of freedom, where k is the number of studies in the meta-analysis and α is usually chosen as 0.05, to give a 5% significance level and thus 95% prediction interval [32].

1.4.3 Individual patient data (IPD) meta-analysis

The meta-analysis of IPD (note: sometimes alternatively called individual participant data) involves obtaining and then synthesising the raw individual level data from multiple related studies [33]. The raw data collected for each individual in a study is regarded as IPD. Aggregate data relates to information which has been summarised or averaged across all individuals in a study, for example the mean treatment effect, proportions of individuals in sub-groups. This aggregated data is derived from the IPD, and the IPD is regarded as the original source [33].

Like the aggregate meta-analysis method, the IPD meta-analysis method aims to synthesise evidence to answer a clinical research question, e.g. a predictive factor being statistically significant. It is important for an IPD meta-analysis to preserve the clustering of patients within studies as it would be inappropriate and introduce possible bias if they were all regarded as being from one study. The preservation of patient clusters can be achieved by using two different approaches [33], the first being a two-step approach [33, 34] where the IPD are analysed separately for each study and aggregate data is then obtained for each study and synthesised using an appropriate meta-analysis method (fixed-effects or random-effects). The one step approach [33] requires simultaneous modelling of the IPD from the studies and can either regard the IPD as coming from one study or account for the cluster of patient within studies. However, there has been very little consideration of one-step and two-step models in risk prediction research [35], which will be discussed in Chapter 3.

1.4.3.1 Why is it potentially useful?

IPD meta-analysis is the gold-standard as there are many advantages over aggregate data method, as the aggregate data are usually poorly presented and reported and may not be available at all in some studies [33]. It can be derived differently for different studies, e.g. some studies may look at odds ratios whilst others may look at the risk difference. Aggregate results that are clinically or statistically significant are more likely to be published, and this causes publication bias [24, 33]. With IPD it is possible to calculate the aggregate data for oneself, so you are not reliant on published reports. IPD allows the assessment of more patients and more outcomes than those that were considered in the original study, which in turn means that IPD meta-analyses are

probably more reliable than aggregate data meta-analyses [33]. IPD also has the ability to identify common patient characteristics within subgroups of patients and tailor the meta-analysis results to them. It also allows the modelling of non-linear effects and time dependent treatment effects, and potentially allows longer follow-up times. Further details about statistical models for IPD meta-analysis are given elsewhere (Stewart *et al.* [36], Simmonds *et al.* [37], Riley *et al.* [38], Higgins *et al.* [39]). Clearly, when developing risk prediction models from multiple studies, it is fundamental to have the IPD available for model development. Further, when evaluating model performance, one might extract aggregate data results from published validation studies; however, if IPD are available one can calculate the performance statistics directly. Thus, IPD would appear preferable for risk prediction research; however this has not been evaluated in detail.

1.4.3.2 Limitations of IPD

There are also a few limitations of IPD. One of the limitations involves trying to obtain the IPD [40], as it may be difficult to contact the authors or trialists as they may have changed institutes or their contact details may be out of date (e.g. e-mail address). By the same token, issues regarding the IPD can be solved through personal contact or travel to ensure the transfer of data, but this may be time-consuming and expensive. Also, researchers may be reluctant to share their data as they have spent a lot of money on the trial, are possessive over their data, or have confidentiality agreements in place [35]. Success may be improved if one has a clear end goal, such as a joint publication in which collaborators are named co-authors.

1.5 Aims and thesis summary

The aim of this thesis is to consider meta-analysis of risk prediction research using both aggregate data and IPD, to gain further insight into these approaches and make methodological applications, evaluations and advancements. In particular, to:

- evaluate the benefit of a meta-analysis of aggregate data for summarising and comparing the performance of risk prediction models, using a real example in breast cancer
- review the methods and reporting used in existing risk prediction research that utilised IPD from multiple studies
- examine the impact of accounting for (versus ignoring) clustering of patients within studies, when using IPD from multiple studies to summarise the performance of a risk prediction tool
- use simulation and IPD from real studies to evaluate a recently proposed method for imputing missing predictor results when given only aggregated data from multiple risk prediction studies

To this end, the research should add important value to the current understanding of how meta-analysis methods, and their application, can facilitate risk prediction research.

An overview of chapters is now given.

1.5.1 Chapter 2

This chapter empirically examines the feasibility of summarising the performance of risk prediction models using a systematic review and meta-analysis of published studies that have developed and/or validated such models. For this purpose a systematic review

is performed and, when possible, a meta-analysis of risk prediction models for the first onset of breast cancer. This work has been published (second author publication) [41].

1.5.2 Chapter 3

The aim of this chapter is to perform a systematic review to assess how risk prediction models are being developed and validated when IPD is sought from multiple studies and combined. The aim is to identify current research techniques and standards; the role of IPD meta-analysis methods towards development and validation; the methodological challenges and problems faces by researchers; and reporting standards. This work has been published (first author publication) [16].

1.5.3 Chapter 4

The aim of this chapter is to illustrate the benefits of having IPD for meta-analysis of risk prediction studies, and to compare how a simple analysis (unstratified) that treats all IPD as coming from a single study compares to a more sophisticated analysis that accounts for clustering. The focus is on a single predictive test and its discrimination ability when using unstratified and meta-analysis approaches. This work has been published (second author publication) [42].

1.5.4 Chapter 5

This chapter extends chapter 4, and aims to examine how the analysis approaches (unstratified, meta-analysis or internal-external cross-validation) impact on the calibration performance of a test for risk prediction and its optimal threshold. This work has been published (second author publication) [42].

1.5.5 Chapter 6

The aim of this chapter is to empirically evaluate the use of a linear imputation method, as suggested by Riley et al. [43], for meta-analysis of test accuracy studies when the thresholds reported by each study differ (where 'threshold' relates to the cut-point used to define positive and negative test results). In each study, the method imputes two by two tables for any missing thresholds that are bounded between two reported thresholds; this enables additional studies to be included in each threshold's meta-analysis. A simulation study is used to compare the true performance of the test (from IPD) to its performance when missing thresholds are generated. This work has been published (second author publication) [43].

1.5.6 Chapter 7

This is the discussion chapter, which will recap the findings from the previous five chapters, discussing their key findings and impact, and outlining limitation and potential extensions.

CHAPTER 2: RISK PREDICTION OF BREAST CANCER: SYSTEMATIC REVIEW AND META- ANALYSIS OF AGGREGATE DATA FROM PUBLISHED STUDIES

2.1 Introduction

This chapter empirically examines the feasibility of using a systematic review and meta-analysis of published studies for the purpose of summarising the performance of risk prediction models. For this purpose a systematic review will be performed and, if possible, a meta-analysis of risk prediction models for the first onset of breast cancer. There are many breast cancer models in the literature, with either single or multiple factors, it is important to identify what these models are telling us. This review will only include models that have a modifiable risk factor, this is important as women are interested in whether they can reduce their risk of first onset of breast cancer [41]. The aims of this systematic review are to identify relevant primary studies that develop or evaluate such a risk prediction model; to qualitatively summarise the content and quantitatively summarise the performance of the models they develop; and to assess if and how the proposed models have been validated. Further, to then consider whether meta-analysis of such primary studies is possible using the aggregate results from study publications. For example, to synthesise reported predictor effect estimates or validation performance statistics across multiple studies of the same prediction model, to reveal

how good these models are and how consistent (or heterogeneous) model parameter estimates are.

This systematic review was performed in conjunction with Dr Catherine Meads, a Senior Lecturer with expertise in systematic reviews at Queen Mary, University of London, and Dr Richard Riley. Dr Meads conceived the project, developed the search strategy and did the literature review. My contributions were to re-check the review classifications from its initial to final phase, to find any discrepancies and resolve them with Dr Meads and Dr Riley; to obtain a final list of relevant articles and extract suitable data (e.g. details about the fitted models, their performance and validation statistics etc.); to appraise the quality of reporting in the articles identified; and to perform meta-analysis where possible. The review has now been published in *Breast Cancer Research and Treatment* (Meads et al. [41]) and is now described in detail, focusing mainly on my contributions.

2.2 Breast cancer

The systematic review focuses on prediction of breast cancer, which refers to a malignant tumour that has developed from cells in the breast. Malignant tumours are cancerous. Left unchecked, malignant cells eventually spread beyond the original tumour to other parts of the body. Usually breast cancer either begins in the cells of the lobules, which are the milk-producing glands, or the ducts, the passages that drain milk from the lobules to the nipple [44].

Breast cancer is the most common cancer in the UK. In 2008, there were 48,034 new cases of breast cancer diagnosed in the UK: 47,693 (over 99%) in women and 341 (less

than 1%) in men. Breast cancer is by far the commonest cancer in women in the UK accounting for 31% of all cases in women. It has been estimated that the lifetime risk of developing breast cancer in 2008 is 1 in 8 for women in the UK [45].

Risk prediction models for breast cancer are very important as they allow assessment of cancer risk in individuals based on their characteristics (predictors). Such models will contain predictors that are deemed clinically and/or statistically significant, and will allow the assessment of risk and prognosis for an individual based on their specific predictor values. These models thus have the potential to save time, money, suffering and death, by identifying in advance those at high risk of developing cancer. Such individuals could then be closely monitored, seek to modify predictor values (and thus reduce cancer risk), and even have early preventative treatment (e.g. breast removal).

2.3 Objectives of the systematic review

The following were the pre-defined objectives of the systematic review:

- 1) To identify and summarise all articles reporting the development and/or the validation of a prediction model for breast cancer, where the model contained at least one modifiable predictor. It was important to look at prediction models containing at least one modifiable risk factor as this can help inform women of how they can reduce their risk of breast cancer and, if altered, to what extent their risk of breast cancer is reduced [41].
- 2) To document the proposed risk prediction models, including the variables in each model and the modifiable predictors included.

- 3) To perform, where possible, meta-analysis of model performance by:
 - i) combining validation statistics for the same model reported in different studies, for both discrimination and calibration performance
 - ii) combining parameter estimates from the same model estimated in different populations (studies), and to examine between-study heterogeneity of parameter values
- 4) To evaluate the reporting quality of the studies identified.

2.4 Methods

2.4.1 Searching and screening

2.4.1.1 Searching for potentially relevant articles

Dr Meads developed a protocol and undertook a search for published articles in November 2009. The databases that were searched were: the Cochrane library, MEDLINE, EMBASE, CAB Abstracts and PsychINFO. The search terms that were used were ‘breast cancer’ and ‘prediction’ or ‘risk model’ as index terms and text words. Reference lists of systematic reviews were also thoroughly checked [41].

2.4.1.2 Inclusion and Exclusion Criteria

The inclusion criteria were studies that either developed and/or validated a breast cancer risk prediction model for first onset of breast cancer in females. In addition, the review focused only on models with at least one *modifiable* predictor in order to be able to inform policies and health decisions. A modifiable predictor is one that can be altered by the individual, i.e. alcohol consumption, BMI, weight or physical activity. For

example, many women are interested in whether they can reduce their risk of breast cancer so it would be useful to know which modifiable predictors are included within existing prediction models. Clinicians and health policy makers may also aim to lower the population rate of breast cancer by public health interventions intended to reduce modifiable predictors. Studies with multiple models are presented. Studies were excluded if they had models with just a single risk factor (it is important to identify models that include a modifiable risk factor as well as other known risk factors, as it is important to know the effect of the modifiable risk factor after it has been adjusted for the other risk factors within the model); had models including only genetic mutations or genes; or predicted anything other than first onset of breast cancer. Early detection, screening and any models published more than 25 years before (i.e. before 1985) were also excluded. This was because of the advancements in medical research in the last 25 years and it would be more useful to assess models that have been developed or validated on data that is reasonably recent.

2.4.1.3 Screening

All identified citations (titles with or without abstracts) were screened by Dr Meads, and then I checked approximately 10% of Dr Meads classifications. Of those that were regarded as potentially relevant, the full articles were ordered and screened further to assess their relevance. The classification was first conducted by Dr Meads and then later checked and modified by myself with a few discrepancies resolved through discussion with Dr Riley. Each relevant article was classified as *developing* a new model, *validating* a model, or a combination of these.

2.4.2 Data extraction

2.4.2.1 Quality of included studies

When the review was undertaken, there were no specific quality assessment checklists for prediction modelling studies. However, a list of key criteria was available by Altman [46]. Based on this paper I developed a quality assessment checklist and extracted the data (shown in Tables 2.1 and 2.3), which was checked by Dr Riley; any disagreements found were resolved via discussion. The quality factors that Altman [46] discusses, which are relevant to prediction models and thus used in this review are as follows:

- Study design – It is advantageous to have a cohort with a long enough follow-up so that enough events are observed (for breast cancer this would be at least five years) for assessment of a number of outcomes (e.g. disease recurrences, disease onset or death)
- Patient sample – a well-defined cohort of patients is required with patient characteristics recorded; eligible patients should not be excluded because of missing data or loss of follow-up
- Sample size – it is important to recognize the power of a study depends on the number of observed events and not the number of patients; a small sample with a longer follow up may be better than a large sample with a short follow up
- Incomplete data, missing or losses to follow up – these are common and serious problem for studies developing a prediction model. This can reduce power and introduce a risk of bias. Completeness of data should be reported by predictor and overall.

- Predictors – It is better to keep continuous variables on their original scale rather than dichotomizing the variables which reduces power and loses important information, as a constant risk is assumed for all the patients in a dichotomized group
- Presentation of multivariate model - whether full presentation of the model is given including all of the predictors, their parameter estimates (predictive factors, i.e. betas (Equations 1.1-1.5)), standard errors or confidence intervals of the parameter estimates, and the baseline risk estimate (alpha or intercept term); these are essential if the model is to be applied in practice
- Validation of the model – the model should be validated; at the very least it should be validated internally which would involve splitting the sample to use a proportion to develop the model and the other part to validate it. Ideally the model should be validated in an external population to give more independent validation results and an indication of generalisability.

For each study in the review, I extracted information relating to each of these quality criteria and noted when important information (e.g. missing data, model parameter estimates, etc.) were not reported. Study characteristics were also extracted and summarised qualitatively. All data extractions were checked by either Dr Riley or Dr Meads.

2.4.2.2 Performance (validation) statistics

For studies aiming to validate a prediction model, I extracted any metric of model performance (such as the E/O statistic and the C statistic (section 1.3.2.2)) alongside their uncertainty (e.g. their confidence interval or standard error); this was checked by

Dr Riley. Where observed over expected (O/E) rates were given in papers, I converted these to E/O statistics to give a consistent scale for meta-analysis. Some studies gave a relevant statistic (e.g. E/O or C) but without the associated standard error. When these situations occurred I needed to indirectly estimate the standard error, to facilitate subsequent meta-analysis. For the C statistic, an equation for the calculation of the CI of the C statistic was used, where C is the C statistic:

$$(C \pm Z_{1-\alpha/2} \text{Standard Error}(C))$$

(Equation 2.1)

Thus if a confidence interval was given, this equation would be rearranged to calculate the standard error. Regarding the E/O statistic, Tice et al. [47] present an equation they used to calculate the 95% CI of the E/O statistic by assuming the observed events follow a Poisson distribution:

$$((E/O) \times \exp(\pm 1.96 \times 1/\sqrt{O}))$$

(Equation 2.2)

It is fine in this instance to use $1/\sqrt{O}$ as the standard error as this event is rare. Thus the standard error of the $\ln(E/O)$ statistic is the inverse of the square root of the observed events, and this was used when the observed number of events was reported; or, if only the confidence interval was reported, equation (2.2) was rearranged to calculate the standard error of the $\ln(E/O)$ statistic when only the confidence interval was given. If a study did not provide variance information, and this could not be derived from other information (such as the confidence interval), then it would not be included in the statistical analysis.

2.4.2.3 Model parameter estimates

For model development studies, the full model estimates were extracted where possible, which included the baseline risk (intercept term), predictors and their parameter (beta) estimates and their uncertainty (confidence intervals or standard errors). Where the full model was not given as the baseline/intercept term was missing, only the coefficients of the predictors in the model and their uncertainty were extracted. If these were not available then the Odds Ratios or Risk Ratios (i.e. $\exp(\beta)$ estimates) were extracted along with their uncertainty, and these were used to obtain the model beta coefficients if possible. In model validation studies, if the model of interest was re-estimated in the new data then the new intercept, parameter estimates and their uncertainty was extracted in the same manner. Extracted results were checked by Dr Riley.

2.4.3 Meta-analysis

2.4.3.1 C statistic and E/O statistic

Where multiple studies reported the same performance statistic for a model, a random-effects meta-analysis was performed, using the DerSimonian and Laird method [48] via STATA 11 [49], to summarise model performance. This approach estimates the average performance of the model, the between-study heterogeneity in model performance, and a 95% prediction interval [32] for the model performance when it is applied in a single population setting. The random-effects meta-analysis framework was introduced in Chapter 1 (Equation 1.15). For the C statistic, the estimates of the C statistic and their standard errors were synthesised. A new paper [27] suggests this is the correct scale to

use for the C statistic. For the E/O statistic, the $\ln(E/O)$ statistics and its standard error was synthesised, and then meta-analysis results were transformed back to the E/O scale.

2.4.3.2 Model parameter estimates

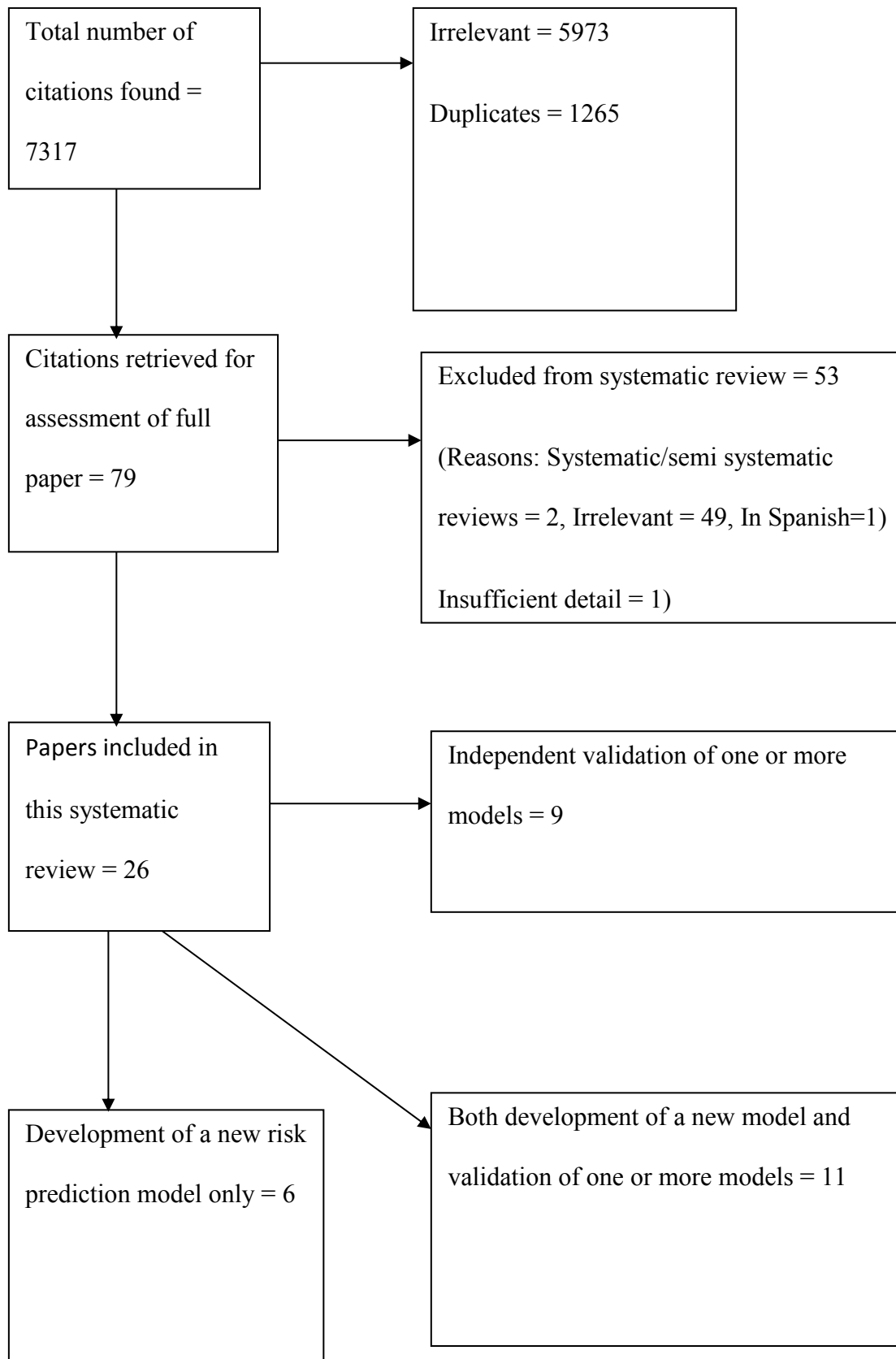
Where the same model was estimated in multiple studies (e.g. in the development study and then in validation studies), the beta coefficient estimate for each predictor were synthesised using a random-effects meta-analysis, again via the DerSimonian and Laird method using STATA 11 [48, 49]. This enabled the effect of each predictor to be summarised and heterogeneity in its effect to be evaluated. This was of particular interest for modifiable predictors, for which the predictive value is of key importance if one considers it to be causal (and thus modifying it can potential reduce breast cancer risk).

2.5 Results

2.5.1 Search results

From the database searches, 7317 references were found of which 1265 were duplicates. The flow of papers toward inclusion or exclusion is shown in Figure 2.1. Dr Meads classified 79 of the 7317 articles as potentially relevant; I went through a random sample of 600 of these classifications (about 10%) and found no discrepancies. However, one article of the 79 retrieved for full assessment was in Spanish and – after discussion with Dr Meads and Dr Riley – it was deemed necessary to exclude this (as we could not translate it), and thus focus only on English language articles. So this left 78 potentially relevant articles, for which the full paper was obtained.

Figure 2.1: PRISMA diagram for systematic review



Dr Meads and I independently went through the 79 papers and further assessed their relevance. A total of 26 articles were deemed relevant by both, with no disagreements, and these were each classed as development and/or validation articles. Of the 26 articles, Dr Meads classified 16 as developing new models and ten as validations of an existing model. In contrast, I classified six as developing a new model, nine as validation of existing model, and 11 as both development of a new model and validation of an existing model. After discussion with Dr Meads and Dr Riley, the classifications by me were accepted and used for the remainder of the review (Figure 2.1).

The 26 articles were grouped as follows:

- Six studies describing the development of a new prediction model (Gjorgov [50], Colditz et al. [51], Cook et al. [52], Gail et al. [53], Tyrer et al. [54] and Wacholder et al. [55])
- Nine studies validating one or more prediction models in a new sample from a potentially different population (Amir et al. [56], Bondy et al. [57], Costantino et al. [58], Rockhill et al. [59], Speigelman et al. [60], Schonfield et al. [61], Ulusoy et al. [62], Rockhill et al. [63] and Viallon et al. [64])
- Eleven studies both describing the development and validation of one or more models (Barlow et al. [65], Boyle et al. [66], Chen et al. [67], Decarli et al. [68], Gail et al. [69], Novotny et al. [70], Rosner et al. [71], Rosner et al. [72], Rosner et al. [73], Tice et al. [74] and Tice et al. [47])

2.5.2 Study quality and reporting

The quality and reporting of the 17 articles developing a risk prediction model is summarised in Table 2.1, and briefly summarised here. None of the 17 papers gave a justification for the sample size used. Fifteen of the 17 papers gave the number of eligible patients for inclusion in to the study for model development (e.g. Barlow et al. [65]) which had index screening mammograms from 1,007,600 patients), but two papers did not, (e.g. Gjorgov [50]). Seven of the 17 gave the number of events per predictor (e.g. Barlow et al. [65]); eight of the 17 (e.g. Boyle et.al [66]) summarised the sample characteristics (e.g. mean age, proportion post-menopausal) in a table; nine of the 17 stated whether there was any missing risk factor data for some participants (e.g. Barlow et al. [65]); and 15 of the 17 stated how they handled continuous variables (i.e. whether kept continuous or categorized) (e.g. Chen et al. [67]). Only six of the 17 reported the full specification of the final developed model(s) (e.g. Cook et al. [52]), i.e. they reported the alpha term with its standard error, as well as parameter values (beta estimates) and their standard errors or 95% confidence intervals for all included variables. Of those 11 papers that did not report the full model, eight of them ignored the intercept but did report beta values or transformed beta values (e.g. odds ratios or risk ratios) for some or all variables, sometimes with a confidence interval (Table 2.1).

2.5.3 Articles developing a prediction model

The statistical model / methods used to obtain the model parameter estimates are shown in Table 2.2 for each of the 17 articles. Nine of those articles used logistic regression, three articles used Poisson regression, two used Cox proportional hazards, two used Bayes' theorem and one article used linear regression. Out of those 17, two have used

Bayes' theorem and one of those articles by Gjorgov [50] has stated: "by calculating exposure to barrier contraceptive practice (condom use and withdrawal practice) along with the factors of parity, age and other (non-barrier) birth-control methods, within a five-year time period and the life span 20-54 years of age, by employing the Bayes' Probability Theorem" [50].

The predictors included in each of the 17 developed risk prediction models are shown in Table 2.3. Note that this does not mean all these gave their full model, as most of these predictors were extracted from tables giving odds ratios or relative risks rather than the full model estimates, as explained above. The modifiable predictors that were included in the models were alcohol consumption, BMI/weight, condom use, exogenous hormone use (HRT, contraceptive pill), and physical activity. The most common predictors included in the models were age, age at first live birth and/or age at subsequent births and family history of breast cancer. The predictors only included in one model were condom use, family history of any cancer, physical activity and reproductive age period. Where condom use was included the justification provided by Gjorgov [50] to include condom use was: "risk-assessment models failed to consider the defined, main, and perhaps the sole most important risk factor and determinant of breast cancer, the exposure to (use of) condoms in marital relations, quantified according to duration ('persistence') of the exposure to condom use (in months and years) during the reproductive-age span of women, from puberty to the peri-menopausal years of 54" [50]. Condom use has only been mentioned by one study, this could be an anomaly and should not be regarded as a predictor solely based on this study [50].

A brief description is now provided of the six risk prediction models developed in these 17 articles.

Gail 1 model [53] used data from a case control study to calculate the parameter estimates for the predictors using logistic regression. These parameter estimates were then combined with the baseline risk estimated using data from the BCDDP Cohort (a population of white women from USA), to allow individualised probabilities to be calculated [41], giving the risk for *any* breast cancer. Gail 2 model is a development of the original Gail 1 model and was published in 1992 as a technical report which is not widely available. This model [69] predicts *invasive* breast cancer only, and the baseline risk is estimated using the SEER database [75]. This means that Gail 1 and Gail 2 have different baseline risks, and indeed are predicting two separate risks.

Tyrer-Cuzick model [54] is based on a Bayesian statistical analysis and was developed in the UK. The population used was from the dataset acquired from the International Breast Intervention Study (IBIS) [76] and UK national statistics. The authors stated: "For an individual woman her family history is used in conjunction with Bayes' theorem to iteratively produce the likelihood of her carrying any genes predisposing to breast cancer, which in turn affects her likelihood of developing breast cancer. This risk was further refined based on the woman's personal history. The model has been incorporated into a computer program that gives a personalised risk estimate."

Pike model [77] is a breast cancer incidence model, based on the observed age-incidence curve and known relations between age at menarche, first birth, and menopause, parity, and the risk of breast cancer [71]. Rosner & Colditz (1) [63] model is a development of the Pike model [77] with an additional predictor allowing for more

than one birth. Rosner & Colditz (2) [73] model is a modification of the Rosner & Colditz (1) [63] model with focus on predicting oestrogen-positive breast cancer. This has been summarised in Table 2.1 below.

Table 2.1 Models and their outcomes

Model	Outcome
<i>Gail 1 model</i>	Any breast cancer
<i>Gail 2 model</i>	Invasive breast cancer only
<i>Tyrer-Cuzick model</i>	Any breast cancer
<i>Pike model</i>	Any breast cancer
<i>Rosner & Colditz (1) model</i>	Any breast cancer
<i>Rosner & Colditz (2) model</i>	Oestrogen-positive breast cancer

Table 2.2: Risk prediction models - reporting and analysis characteristics

Article	Description of key aspects of study design:		Sample characteristics: Were they summarised in a table (e.g. mean age, proportion males, ...)	Data quality: Missing data for each variable mentioned/ stated in the paper?	Handling of continuous variables: Were they kept continuous or categorized in the model?	Presentation of model:		
	Number of eligible patients given	Number of events per variable				Was the complete model given (i.e. alpha term, parameter estimates and their uncertainty, i.e. s.e. or CI)?	If not, was any part of the model given?	What statistical model / method was used?
Gjorgov 2009 Barlow 2006	No Yes, 1007600	Not stated Stated	No Yes	No Yes	Not given Categorised (test for trend across categories of a continuous factor performed)	No No	No Yes – ORs and CIs for each category of each variable relative to the reference category	Bayes' theorem Logistic regression
Boyle 2004	Yes, 5157	Not stated	Yes	No	Categorised	No; (only some of the final model variables are given)	Yes; ORs and CIs given for a partial set of the included variables; score chart provided	Logistic regression
Chen 2006	Yes, 284780	Not stated	No	Yes	Categorised	No	Yes (variable names with coefficients, but no CIs)	Logistic regression
Colditz 2000	Yes, 58520	Not stated	No	No	Kept continuous	Yes		Poisson regression
Cook 2009	Yes, 45281	Not stated	Yes	Yes, as complete data available	Kept continuous	Yes		Logistic regression
Decarli 2006	Yes, 5157	Stated	Yes	Yes – patients with missing data excluded	Categorised	No	Yes; ORs and CI given for variables	Logistic regression
Gail 1989	Yes, 5998	Stated	No	No	All Continuous except age (categorized)	Yes (estimates and their standard errors given)		Logistic regression
Gail 2007	Yes, 3254	Stated	Yes	Yes, as complete data available)	Categorised	Yes (estimates and their standard errors given)	No	Logistic regression
Novotny 2006	Yes, 4598	Not stated	No	No	Categorised	No	Yes; ORs and parameter estimates given, but no standard errors or CIs	Logistic regression
Rosner 1994	Yes, 91523	Not stated	No	Yes, patients with missing data excluded	Continuous	No	Yes – some parameter estimates and some CIs given	Poisson regression
Rosner 1996	Yes, 89132	Not stated	No	No	Continuous	Yes		Poisson regression
Rosner 2008	Yes, 59812	Not stated	No	No	Continuous	Yes		Linear regression

Article	Description of key aspects of study design:		Sample characteristics: Were they summarised in a table (e.g. mean age, proportion males, ...)	Data quality: Missing data for each variable mentioned/ stated in the paper?	Handling of continuous variables: Were they kept continuous or categorized in the model?	Presentation of model:		
	Number of eligible patients given	Number of events per variable				Was the complete model given (i.e. alpha term, parameter estimates and their uncertainty, i.e. s.e. or CI)?	If not, was any part of the model given?	What statistical model / method was used?
Tice 2005	Yes, 81777	Stated	Yes	Yes	Categorised, as in the Gail model	No	Yes – some parameter estimates and CIs given	Cox proportional hazards
Tice 2008	Yes, 1095484	Stated	Yes	Yes	A mixture of continuous and categorisation used	No	No	Proportional hazards model (probably cox but not stated)
Tyrer 2004	No	Not stated	No	No	Not given	No	No	Bayes' theorem
Wacholder 2010	Yes, 11588	Stated	Yes	Yes	Categorised	No	Yes, ORs and CIs for the variables	Logistic regression

Table 2.3: Breast cancer risk prediction models - comparison of predictors included

Predictors	Gjorgov 2009	Barlow 2006 pre- menopausal	Barlow 2006 post- menopausal	Boyle 2004	Chen 2006	Colditz 2000	Cook 2009	Decarli 2006	Gail 1989	Gail 2007	Novotny 2006	Rosner 1994	Rosner & Colditz 1996	Rosner 2008	Tice 2005	Tice 2008	Tyrer- Cuzick 2004	Wacholder 2010 models**
Age	Y	Y	Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	1,2,3
Age at menarche				Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y	1,2,3
Age at first live birth and/or age at subsequent births			Y	Y	Y	Y	Y		Y	Y	Y	Y	Y	Y	Y		Y	1,2,3
Age at menopause						Y	Y					Y	Y	Y			Y	
Atypical hyperplasia / benign breast disease			Y			Y	y				Y		Y	Y			Y	
Breast density		Y	Y		Y										Y	Y		
Birth history/ parity	Y					Y								Y			Y	
Birth index						Y	Y											
Breast biopsy number		Y	Y		Y			Y	Y	Y	Y		Y		Y	Y		1,2,3
Ethnicity			Y													Y		
Family history of breast cancer	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y			Y	Y	Y		1,2,3
Family history of any cancer											Y							
Height						Y	Y							Y			Y	
Number of contraceptives											Y							
Reproductive age period	Y																	
Surgical menopause			Y			Y	Y							Y				
Modifiable Predictors																		
Alcohol consumption				Y		Y	Y							Y				
BMI or weight			Y	Y	Y	Y	Y				Y			Y			Y	
Condom use	Y																	
Exogenous hormone use (pill, HRT)			Y	Y		Y	Y							Y				
Physical activity				Y														
**Five models were presented in Wacholder 2010 but only three had modifiable predictor and have been presented here																		
1,2,3 represent the three Wacholder 2010 models																		
A total of 21 predictors have been identified																		

2.5.4 Articles externally validating a model

2.5.4.1 Qualitative summary

The characteristics and results of the 20 articles externally validating one or more prediction models are shown in Table 2.4. These are now summarised qualitatively.

There were four papers with external validations of the Gail 1 model (Bondy et al. [57], Costantino et al. [58], Novotny et al. [70], Spiegelman et al. [60]) and 12 papers with external validations of the Gail 2 model (Amir et al. [56], Barlow et al. [65], Boyle et al. [66], Chen et al. [67], Costantino et al. [58], Decarli et al. [68], Gail et al. [69], Rockhill et al. [59], Schonfeld et al. [61], Tice et al. [74], Tice et al. [47], Ulusoy et al. [62]).

Costantino et al. [58] evaluated both Gail models, and several other papers presented more than one validation using different cohorts. The validation populations were generally from USA, but there was also one study from each of Czech Republic, France, Great Britain, Italy and Turkey.

None of the four papers assessing Gail model 1 reported a C statistic, but three of them reported an E/O statistic. Nine of the 12 papers that assessed Gail model 2 gave the C statistic, while seven of the 12 papers gave the E/O ratios. Five of these papers [58, 59, 61, 69, 74] refitted the original model and gave the new parameter estimates and their uncertainty.

Two papers [63, 64] assessed the Rosner and Colditz (1) model and two papers [63, 73] assessed the Rosner and Colditz (2) model; one paper from these has assessed both models. Two papers gave the C statistic for Rosner and Colditz (2) model, and one paper gave the C statistic for Rosner and Colditz (1) model. Two papers gave the E/O

ratio for Rosner and Colditz (1) model, and one gave the E/O ratio for Rosner and Colditz (2) model. Only one paper refitted the Rosner and Colditz (2) model in a new data set and gave the parameter estimates and their uncertainty.

Two papers [71, 72] assessed the Pike model but neither gave the C statistic or the E/O ratio.

One study (Amir et al. [56]) (Table 2.4) was found which assessed the performance of several models (Gail, Claus, Ford, Tyrer-Cuzick and Manual) on the same data set from South Manchester, UK. The population of 4536 women had been assessed in a hospital clinic for breast and other cancer risks and were a high risk sample. Amir et al. [56] investigated the total population and the screened population, and they concluded that the Tyrer-Cuzick model was the most accurate for this high risk sample with a C statistic of 0.762 (0.700-0.824) (see Table 2.4). It is not clear whether the Tyrer-Cuzick model would also be the best predictive model for a general population sample (a general setting), but this study is helpful as it gives an insight as to how some of the proposed risk prediction models perform on the same dataset. Tyrer-Cuzick has a better C statistic value than the other models, and therefore it appears to discriminate better than the other models.

Table 2.4: Breast cancer risk prediction models - validation statistics

Article	Model validated	C Statistic (95%CI)	Other validation stats	Validation population	Comments
Amir 2003	Gail 2	0.735 (0.666 to 0.803)	E/O = 0.69 (0.54 to 0.90)	High risk hospital cases and controls from UK	Also evaluated three genetic models
	Tyrer-Cuzick	0.762 (0.700 to 0.824)	E/O = 1.09 (0.85 to 1.41)		
Barlow 2006	Gail 2	0.598 (CI and SE not given)	Not given	Mammogram registry in USA population	Differences in time intervals for cancer ascertainment meant validation unreliable
Bondy 1994	Gail 1		O/E = 0.76 (CI not given, but calculated for meta-analysis when converted to E/O and using Equation 2.2)	High risk white women in Texas USA	Subgroup analysis based on American Cancer Society mammogram screening guidelines given
Boyle 2004	Gail 2	Not given	O/E = 0.89 (0.70 to 1.09)	RCT of adjuvant tamoxifen in USA, all had hysterectomies	2 validations, above with original USA data set, below with Italian registry dataset. Unclear if for pre- or post-menopausal women or both.
		0.582 (CI and SE not given)	O/E = 0.96 (0.75 to 1.16)		
Chen 2006	Gail 2	0.602 (CI and SE not given)	Not given	Unclear	
Costantino 1999	Gail 1	Not given	E/O = 0.84 (0.73 to 0.97)	Women at increased risk of breast cancer in USA RCT of adjuvant tamoxifen	Distinguished clearly between total breast cancer (Gail 1) and invasive breast cancer (Gail 2)
	Gail 2		E/O = 1.03 (0.88 to 1.21)		
Decarli 2006	Gail 2	0.588 (0.546 to 0.631)	E/O = 0.93 (0.81 to 1.08)	Italian case control and registry studies	-
Gail 2007	Gail 2	0.636 (0.617 to 0.655)	O/E = 1.08 (0.97 to 1.20)	Black women from USA	Recalculated c statistic and O/E from data in paper
Novotny 2006	Gail 1	Not given	Not given	Mammogram registry in Czech population	Only parameter estimates (ORs) with no standard errors of CIs given
Rockhill 2001	Gail 2		E/O = 0.94 (0.89 to 0.99)	White nurses in USA	Subgroup analyses for high risk and mammogram in past year also given
Rockhill 2003	Rosner & Colditz (1)	0.57 (0.55, 0.59)	E/O = 1.00 (0.93 to 1.07)	Nurses from USA	Validation on same cohort as original model but using different time ranges
	Rosner & Colditz (2)	0.63 (0.61, 0.65)	E/O = 1.01 (0.94 to 1.09)		
Rosner 1994	Pike	Not given	Not given	Nurses from USA	-
Rosner 1996	Pike	Not given	Not given	Nurses from USA	-
Schonfield 2010	Gail 2	Not given	Early SEER E/O = 0.87 (0.85 to 0.89)	White postmenopausal women from USA (NIH-AARP study)	Split SEER cohort by date from 1983-87 and 1995-2003 and validated using two different populations.
			Late SEER E/O = 1.03 (1.00 to 1.05)		
Rosner 2008	Rosner & Colditz (2)	0.635 (0.628 to 0.642)	Early SEER E/O = 0.86 (0.82 to 0.90)	White postmenopausal women from USA (PCLO trial)	Focus of paper on oestrogen receptor-positive breast cancer
			Late SEER E/O = 1.01 (0.97 to 1.06)	Nurses from USA	

Article	Model validated	C Statistic (95%CI)	Other validation stats	Validation population	Comments
Speigelman 1994	Gail 1	Not given	E/O = 1.33 (1.28 to 1.39)	Nurses from USA	Over prediction attributed to higher baseline incidence rates of breast cancer
Tice 2005	Gail 2	0.67 (0.65 to 0.68)	Not given	Mammography register in USA	ROC curve symmetrical
Tice 2008	Gail 2	0.613 (0.604 to 0.622)	Not given	7 mammogram registries from USA	Some data missing so authors recommend interpretation with caution
Ulusoy 2010	Gail 2	Not given	Using cut off risk ≥ 1.67 sensitivity = 13.3%, specificity = 92%, PPV = 63%, NPV = 51.9%	Turkish cases and controls from one hospital	Small validation sample
Viallon 2009	Rosner & Colditz (1)	Not given	E/O = 0.947 (0.912 to 0.982)	French teacher, spouses and employees	Most of paper describes mathematical simulations.

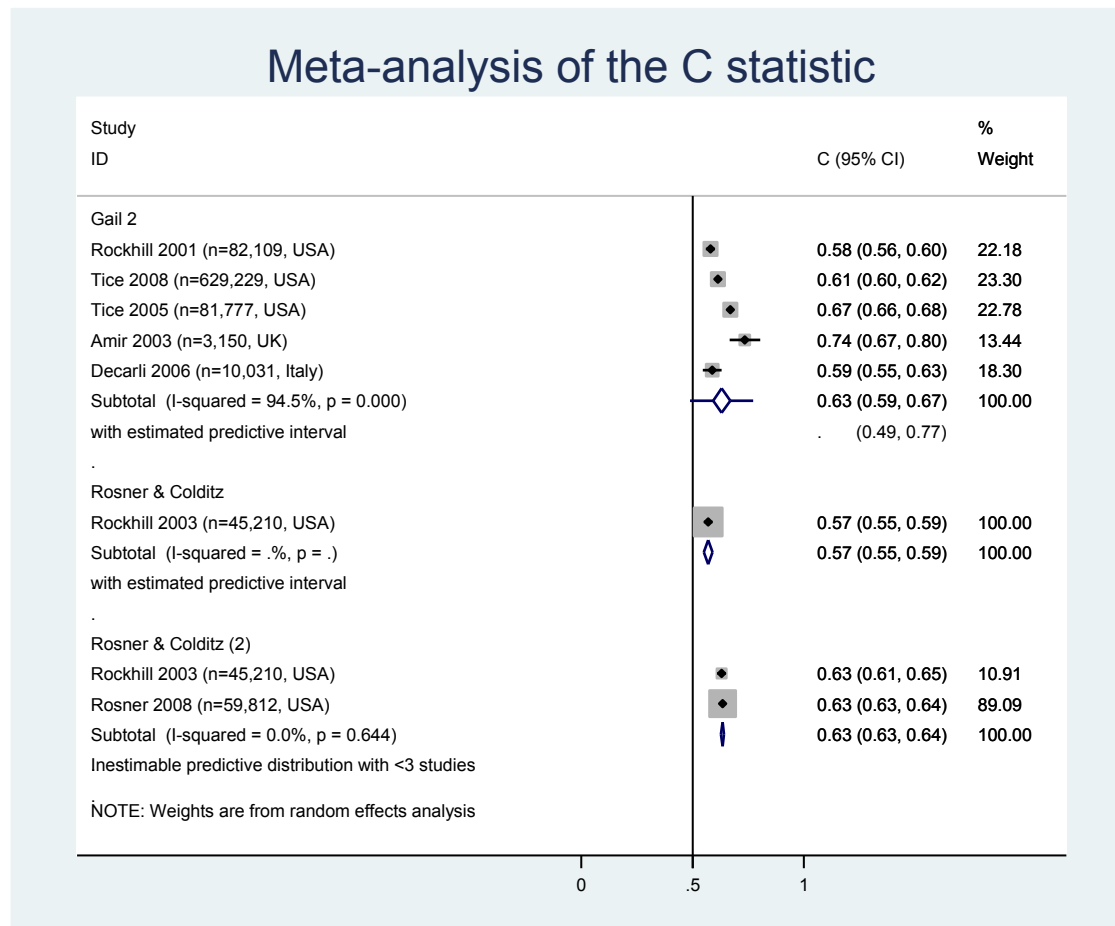
2.5.4.2 Meta-analysis of external validation statistics

Random-effects meta-analyses of the C Statistic (Figure 2.2) and the E/O ratio (Figure 2.3) were conducted for each model that had the external validation statistic extracted from two or more different articles (or populations). Validation statistics were often not reported, thus only a few meta-analyses were possible which usually contained a small number of studies. In most analyses there was considerable between-study heterogeneity in the validation statistic; for example, for the Gail 2 model the proportion of the total variability due to between-study heterogeneity (I^2) was 94.5% and 92.5% in the meta-analyses of the C statistic and E/O ratio respectively. Such heterogeneity is perhaps unsurprising given the variations in populations used for the validations (Table 2.4). However, I-squared is also large because the within-study variances are small (due to the number of patients being extremely large in each study), and so the between-study variance (tau-squared) always dominates. Rucker et al. [78] refers to this as a ‘misleading’ nature of the I^2 statistic when the within-study variances are very small.

It was only possible to do a meta-analysis of the C statistic for Gail 2 and Rosner & Colditz (2), for which only five and two studies gave results (Figure 2.2). Rosner & Colditz (1) is included in Figure 2.2 as that was the only other article to give a C statistic with its uncertainty, so it was included so all available data can be viewed on the graph. It is difficult to make strong inferences. In terms of discriminatory performance of the models, generally the C statistics across studies seem to be around 0.6, and thus only moderate; but nevertheless, they are adding some potentially important information about increase in breast cancer risk when compared to a null model with no predictors (i.e. where the C-statistic would be 0.5, a 50:50 (tossing a

coin) prediction). No result contains 0.5 in the 95% CI, so it seems these models can discriminate between those women who develop breast cancer and those who do not, but not too well.

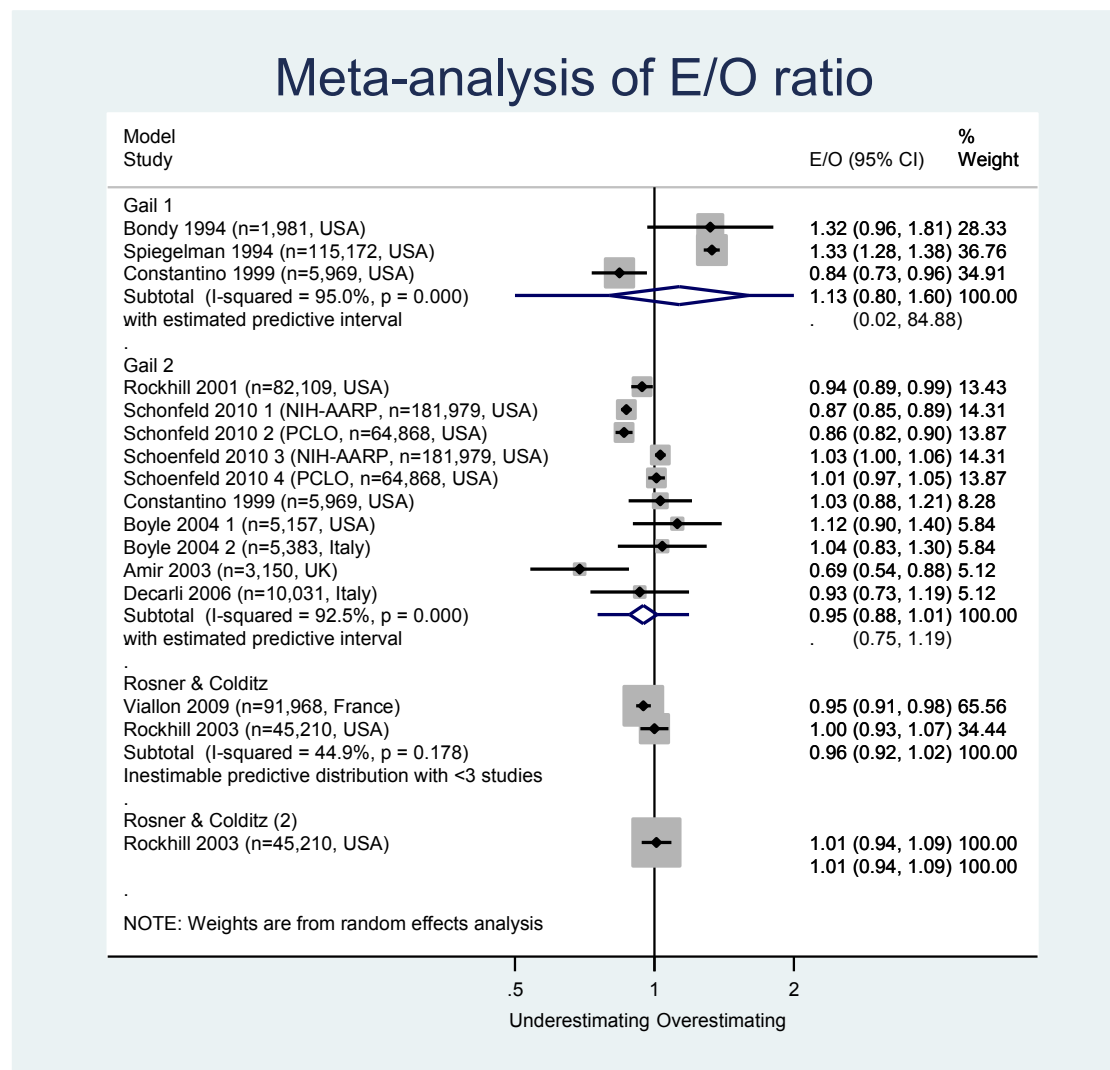
Figure 2.2: Meta-analysis of the C statistic for breast cancer risk prediction models



For meta-analyses of the Gail 2 model, the average C statistic was 0.63 (95% CI: 0.59 to 0.67, based on 5 articles) which indicates only moderate performance. Gail 2 has a 95% prediction interval of (0.49, 0.77), which indicates that (based on the 5 studies in the meta-analysis) we predict that the C statistic in a new study (population) would be between 0.49 and 0.77. This reveals that the potential C statistic in a particular clinical

setting may be substantially different from the average effect [32], with the 95% prediction interval indicating a wide-range of patient values, including even 0.50, and therefore it is unclear whether the model will have adequate discrimination in a particular setting of application. One can formally check if the model results are actually different if all the models were applied to the same data and then the C statistics were to be compared, i.e. Amir et al. [56] who do this for Gail 2 model and Tyrer and Cuzick model. This shows how well the models perform against each other in the same setting.

Figure 2.3: Meta-analysis of the E/O ratio for breast cancer risk prediction models



In terms of calibration of the models, the meta-analysis of E/O statistics is shown in Figure 2.3. Such a meta-analysis was only possible for Gail 1, Gail 2, and Rosner & Colditz (1) as the results were presented for three, seven and two studies respectively. Rosner & Colditz (2) is included in Figure 2.3 as that was the only other article to give an E/O statistic with its uncertainty so it was included so all available data can be viewed on the graph. There appears to be a high level of heterogeneity in the meta-analyses for Gail model 1 (95%) and Gail model 2 (92.5%), again likely to be due to the range of different populations considered and the ‘misleading’ nature of I^2 given small within study variances (alternatively tau-squared can be used, as it is less sensitive to the sample size [78]).

For Gail model 2, this heterogeneity is very important to acknowledge. On average the model has good calibration, with the summary E/O statistic close to 1 and its confidence interval contains 1 (pooled E/O = 0.95; 95% CI: 0.88 to 1.01). However it can be seen in Figure 2.3 that in individual studies E/O is sometimes less than 1 and in others it is greater than 1; on average it is close to 1 and the 95% CI contains 1, this shows on average it is a good model – but there is considerable heterogeneity in calibration across studies. This is revealed by the 95% prediction interval of 0.75 to 1.19 which gives the range of possible E/O ratios when the Gail model 2 is applied in any given study population. On average it performs well but it may not do as well in an individual clinical setting, potentially underestimating the number of events (0.75) or overestimating them (1.19). This is similar for Gail model 1; although the prediction interval is extremely wide given there are only 3 studies in the meta-analysis (Figure 2.3). The meta-analysis of the C statistic in Figure 2 shows the discrimination is quite

low, but the calibration for those similar studies is much better. For a model to be considered good it should have good discrimination (which depends on the clinical context, but ideally it should have a C statistic somewhat greater than 0.50) and calibration (e.g. being as close to 1 as possible for E/O) [13].

2.5.4.3 Meta-analysis of the refitted parameter estimates

In validation studies that refitted the original model under investigation, the new parameter estimates and their 95% confidence intervals or standard errors (uncertainty) were directly or indirectly available for just 5 articles. All 5 articles refitted the Gail 2 model, which contains seven different variables: Age at menarche, Number of biopsies*Age categorised interaction variable (No biopsies <50, No biopsies \geq 50), and four interaction terms for number of affected first degree relatives*age at first birth*with age in four different categories. A separate random effects meta-analysis of the parameter estimates for each of the variables is presented in Figures 2.4 to 2.10. This is another approach of checking the consistency of model parameter estimates across different settings: if there is little heterogeneity and similar direction and magnitude of effects, it enhances the credibility of the model's generalisability. The motivation for this meta-analysis is to assess whether these predictors show consistent predictive value when pooled together across different studies. Ideally, parameter estimates would be available for the modifiable risk factors, so that they can be meta-analysed and the obtained summary results would help ascertain whether these modifiable risk factors (consistently) reduce the risk of breast cancer in all populations.

Figures 2.4-2.10 show that the parameter estimates are similar across studies in terms of direction, although the heterogeneity is sometimes a concern, ranging from an I-squared of 0% (Figure 2.10) to 84.2% (Figure 2.5).

The summary meta-analysis results reveal that, at least on average, each parameter is a predictor of breast cancer risk, and so there is strong evidence that these are genuinely important predictors to consider in breast cancer risk prediction models. Age at menarche is a predictive factor as shown by its pooled result of 1.08 (1.02, 1.14). The number of biopsies is contained within some of the parameters in this model as an interaction term with age (Figures 2.5-2.6). Comparing the summary results of Figure 2.5 (one biopsy \times age <50) and 2.6 (one biopsy \times age ≥ 50), the summary result is very similar and so there is no clear evidence that age categorised is important (a different threshold for age may provide different results), but both meta-analysis results do suggest that having had one biopsy is a predictive factor.

Figure 2.4: Meta-analysis of age at menarche for Gail 2 models

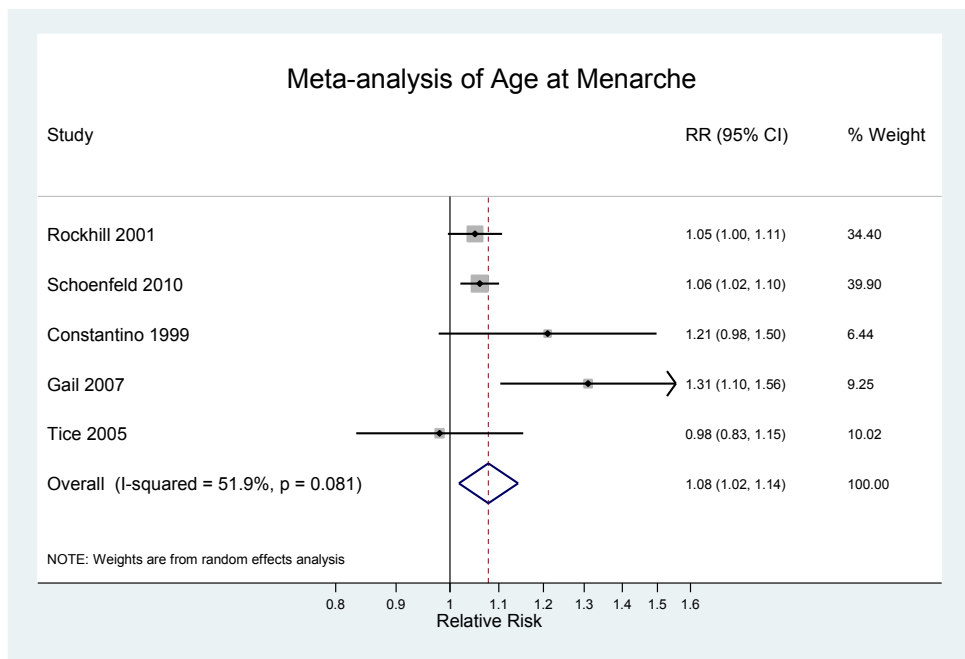


Figure 2.5: Meta-analysis of one biopsy × age categorised interaction term (age <50) for Gail 2 models

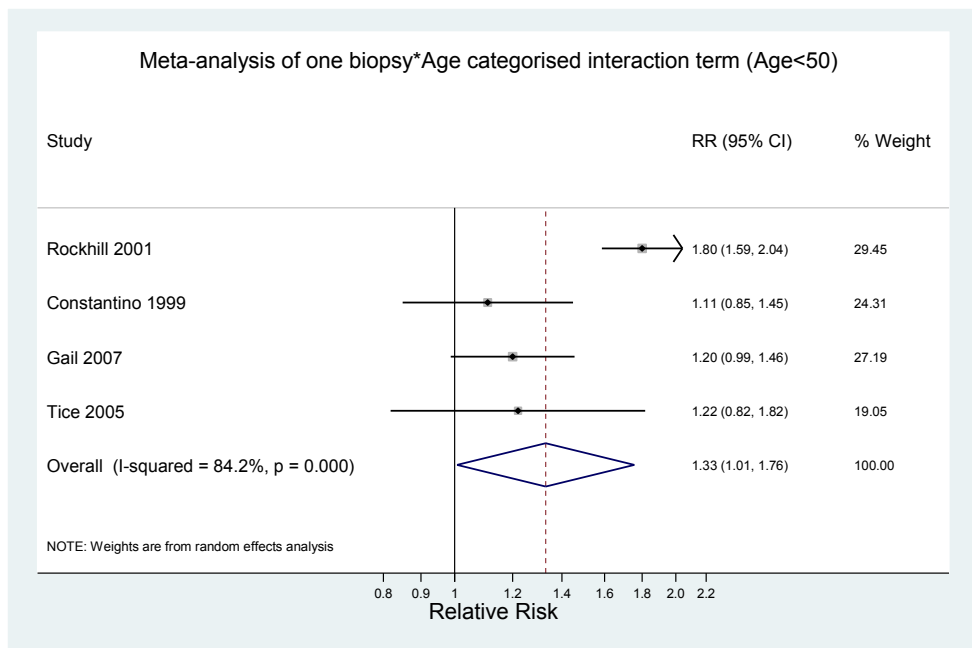


Figure 2.6: Meta-analysis of one biopsy × age categorised interaction term (age ≥50) for Gail 2 models

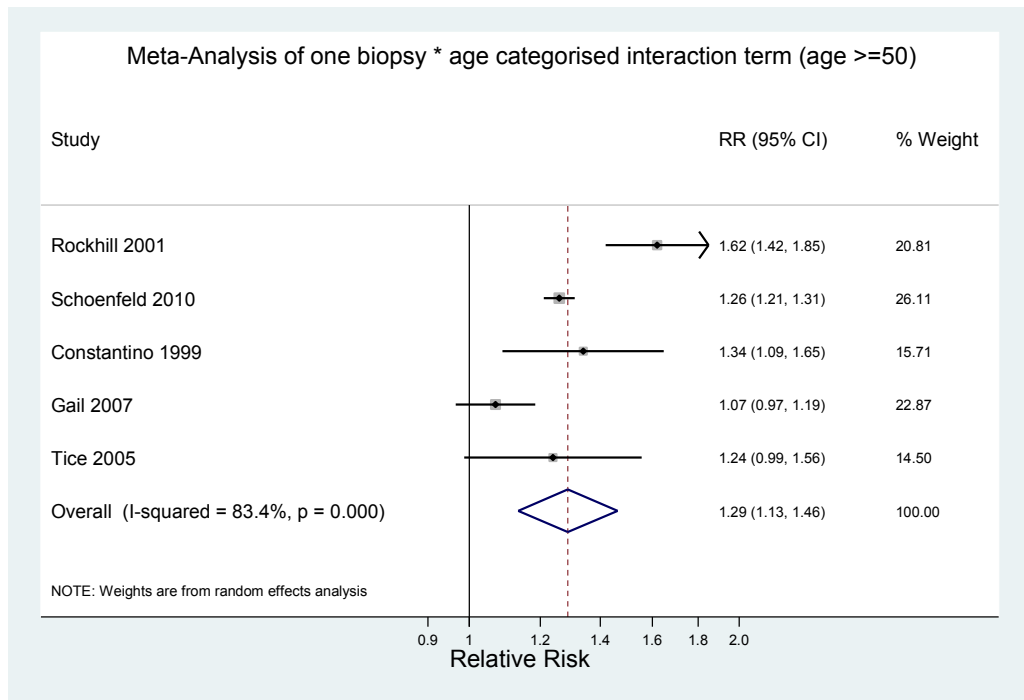


Figure 2.7: Meta-analysis of one affected first degree relatives \times age at first live birth \times age <20 interaction term for Gail 2 models

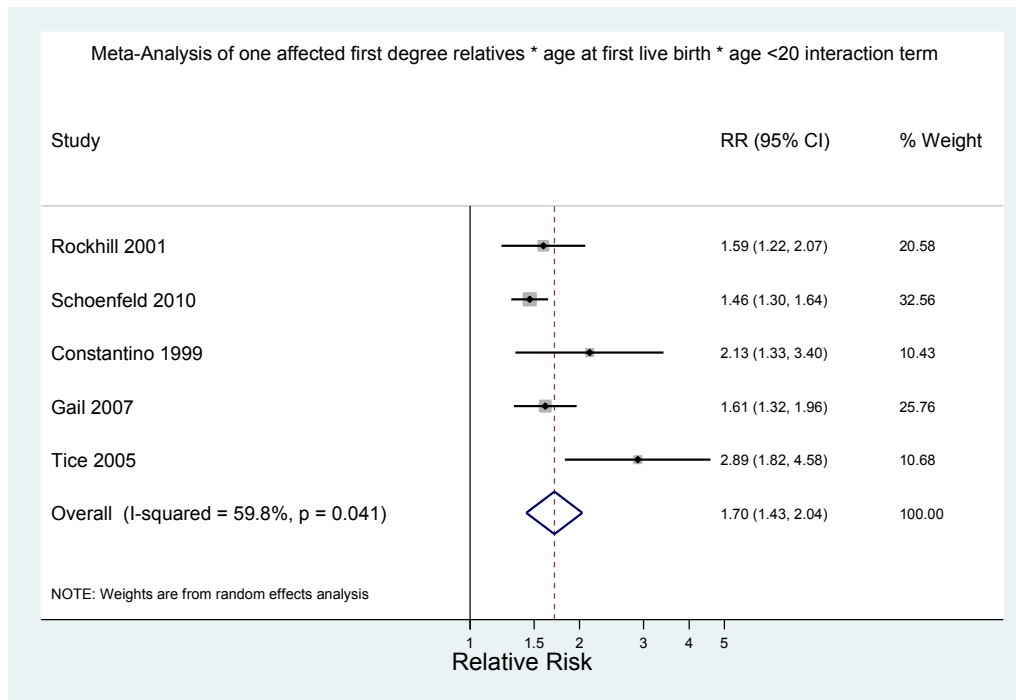


Figure 2.8: Meta-analysis of one affected first degree relatives \times age at first live birth \times age 20-24 interaction term for Gail 2 models

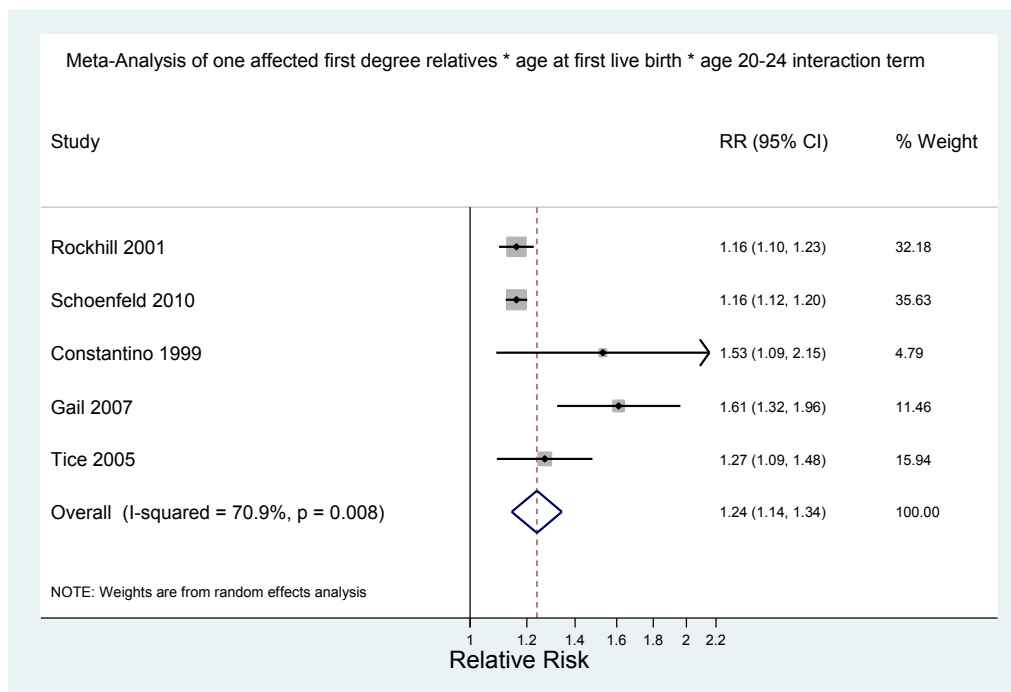


Figure 2.9: Meta-analysis of one affected first degree relatives × age at first live birth × age 25-29 interaction term for Gail 2 models

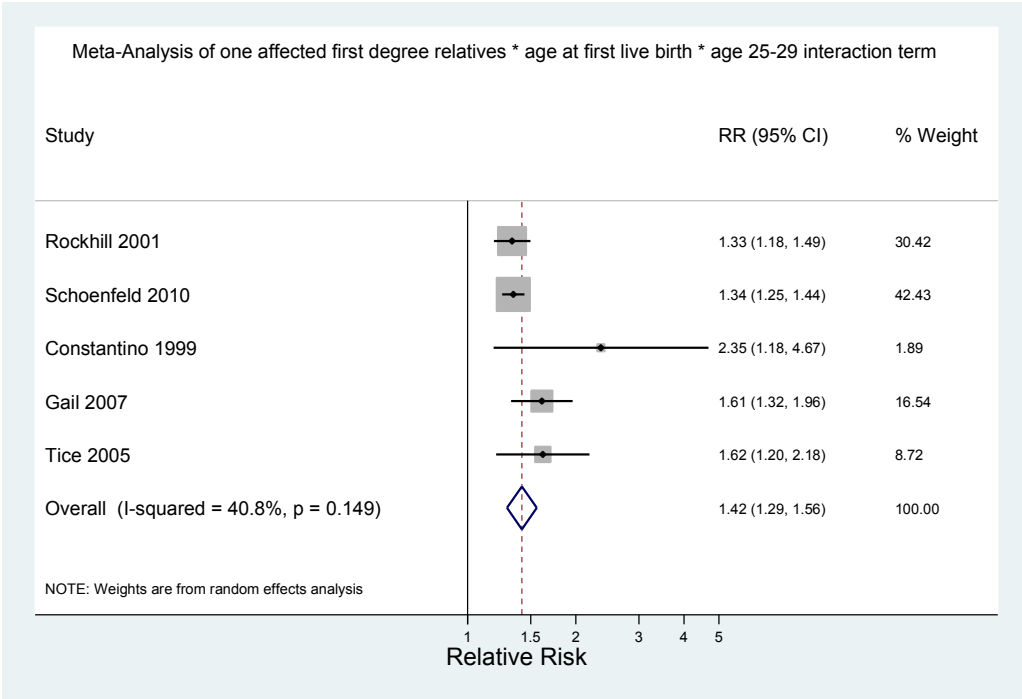
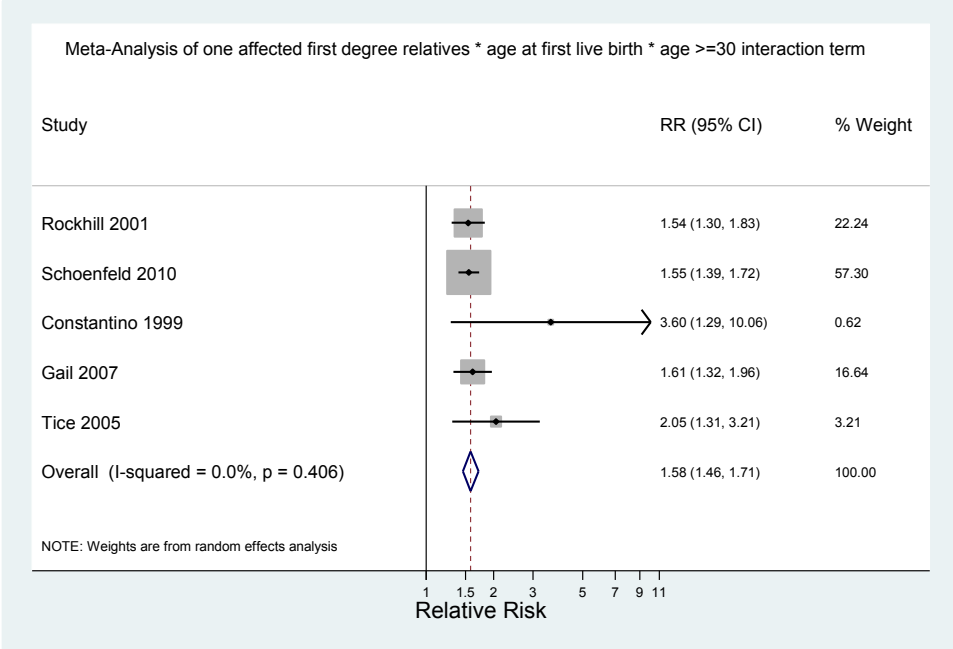


Figure 2.10: Meta-analysis of one affected first degree relatives × age at first live birth × age ≥30 interaction term for Gail 2 models



Figures 2.7–2.10 show the meta-analysis results for interaction terms of one affected first degree relative with age at first live birth and age categorised. Figure 2.7 shows a

pooled result of 1.70 (1.43, 2.04) for age<20; for age 20-24 (Figure 2.8) the pooled result is 1.24 (1.14, 1.34) which rises to 1.42 (1.29, 1.56) (Figure 2.9, age 25-29) and 1.58 (1.46, 1.71) (Figure 2.10, age \geq 30). Figure 2.4 shows that as age increases the risk increases. The interactions here show how family history plays an important role in the risk in those aged <20 and >30.

2.6 Discussion

This chapter has described a systematic review which identified and then summarised proposed prediction models for risk of breast cancer, to clarify the models available, their predictive performance, reporting standards, and contained predictors. The systematic review found 7317 citations, which subsequently resulted in 26 papers being chosen for inclusion in this review. Of these, six developed a new risk prediction model only, nine independently validated one or more models and 11 both developed a new model and validated one or more models. The evaluation of study quality and reporting for papers that developed a model was found to be quite poor in general, as only six of the 17 reported the full specification of the final developed model(s) and the uncertainty for all included variables. The modifiable risk factors found in the models were alcohol consumption, BMI/weight, condom use, exogenous hormone use (HRT, contraceptive pill) and physical activity. A total of 21 different risk factors were included in all the models.

A meta-analysis of C statistics was performed to compare model performance through discrimination, with only five studies providing C statistics along with their uncertainty and two studies providing these for Rosner & Colditz (2). The pooled values for the C statistic for Gail2 was 0.63 (0.59, 0.67) and for Rosner & Colditz (2) was 0.63 (0.63,

0.64). As the 95% CI does not contain 0.5 (which is the null value) these C statistic values indicate the models do provide some moderate discrimination between those who will and will not develop breast cancer, and may thus be potentially useful to identify those at higher risk. A similar meta-analysis was performed for the E/O ratio, with three studies providing data for Gail 1 model (pooled result of 1.13 (0.80, 1.60)), six studies (10 models) providing data for Gail 2 model (pooled result of 0.95 (0.88, 1.01)) and two studies providing data for Rosner & Colditz (1) model (pooled result of 0.96, (0.92, 1.02)). These pooled results for Gail 2 and Rosner & Colditz (1) are close to 1 with the 95% CI being quite narrow and containing 1, suggesting that these models calibrate almost perfectly over all patients. This meta-analysis examined the average performance of particular models, and quantified the heterogeneity in model performance across settings. Further, meta-analyses of predictor estimates to check for consistency of parameter estimates across different settings found that they are still significant parameters in the populations examined.

2.6.1 Limitations and strengths of the review

There are a few limitations to this review. For example only 10 % of the initial reference list was checked by a second reviewer; if the entire reference list was checked there is a possibility more articles could have been identified. Another limitation was only including models with modifiable predictors, which reduced the number of studies that could be included in the review. Also, the meta-analyses and evaluation were limited by poor reporting of the identified studies: in particular, a number of papers did not report the model with its parameters and uncertainty. The model was generally presented in tables as either odds ratios or relative risk ratios, but many of the papers did

not provide the C statistic or E/O ratio, or if they did present those, few also provided their uncertainty. Only the papers that did provide these values with their uncertainty could be used in the meta-analysis.

On a positive note a large data extraction was performed during this review and the results were synthesised as far as possible, to give a quantitative summary of the state of the field. Meta-analysis was possible for some models where they had either the C statistic, E/O ratio or predictor estimate along with its uncertainty. The studies were also qualitatively assessed using a checklist which provided especially useful for evaluating the quality in reporting.

2.6.2 The best model for predicting breast cancer risk?

The review identified six different models proposed for predicting breast cancer risk; the most widely assessed models were Gail models 1 and 2. Gail model 1 and Gail model 2 have different baseline risks, and are predicting two separate risks (any breast cancer versus invasive breast cancer).

From the findings of this review it is very difficult to determine the best model. It was not possible to compare most of the models, as the non-Gail model articles often did not present validation results. Further, the two Gail models both predict two different risks, and thus are not directly comparable. The C statistics are generally low for all models, which means that the models do not discriminate very well between those who will and those who will not develop breast cancer. The meta-analysis of the E/O ratio helps reveal the performance of these models in terms of calibration of observed and model expected events. Gail model 2 on average performs well, which shows on average these

models are predicting risk well in populations; however, the prediction interval for E/O in a new population shows this may not be the case in an individual setting, as they are very wide and contain values far from 1. Prediction intervals have been developed and used for disseminating random-effects meta-analyses of treatment effect studies [32, 79], but here they have been used to help evaluate model performance. As seen in Figure 3, meta-analysis of the E/O ratio gives a 95% prediction interval for Gail 1 which is very wide, due to the small numbers of studies and the observed heterogeneity. The prediction interval for Gail 2 is much narrower than the prediction interval for Gail 1, but there is still considerable potential variability in model calibration performance across settings, and importantly it may not be close to 1 always. Thus it is difficult to conclude that Gail 1 or Gail 2 models will give reliable enough predictions for use in practice: the predictive performance appears to depend on the setting of application, and discrimination only appears moderate.

The meta-analysis of the parameter estimates in the Gail 2 model confirms that the included predictors are likely to be important, as the summary risk ratios in the validation datasets are all in support of an association between the predictors and the risk of breast cancer. Future risk prediction models in this field should consider starting with these predictors, therefore.

2.6.3 Reporting standards

The quality and reporting of the 17 articles developing a risk prediction model is summarized in Table 2.2. The quality of reporting was very poor, which again makes it hard to determine the best model to predict the first onset of breast cancer.

Altman [46] has investigated the reporting standard for papers that propose a prediction model. Although his findings are from studies using prognostic models (start point is disease), the issues are very similar and applicable to studies that generally use risk prediction models (in healthy or diseased people), as found here. The article by Mallet et al. [80] has similar findings as the Altman paper [46], they state: “Many published prognostic models have been developed using poor methods and many with poor reporting, both of which compromise the reliability and clinical relevance of models, prognostic indices and risk groups derived from them.”

Our review found that the description of key aspects of study designs was very poor, patient characteristics were poorly reported and data completeness was poorly reported. Also, of concern was the handling of continuous predictors as a number of studies categorized some or all of the continuous predictors. Of critical importance is the finding that many authors did not even present the prediction model in full, but instead chose to present risk groups derived from the model or some of the parameter estimates but without the intercept (baseline risk).

2.6.4 Next steps for prediction model research in breast cancer

It is imperative from this review that future risk prediction model papers improve the quality of their reporting, and for the models used to be presented in their entirety (with intercept terms and all variables along with standard errors or 95% confidence intervals). For those papers that externally validate it is important for them to report validation statistics (e.g. a C statistic and E/O ratio, with confidence intervals) that will help the reader understand how well the validation has performed and increase/decrease confidence in the model accordingly. It is also important that when these models are

developed they are not only validated on the same dataset they were developed on (internal validation), but they validate it externally on a different dataset, which will show how well it performs in a different clinical setting.

The TRIPOD statement [81] was published after this work was completed, but it echoes similar thoughts. It is a very important publication in prognostic research, as it provides a checklist of 22 items deemed essential for transparent reporting of prediction model studies [81]. If this checklist is adopted by the wider research community, it can help to minimise poor reporting in primary studies. Some of the recommendations include the full model being presented, how to use the model, its performance measures along with 95% CIs, as well as discussing the potential use of the model and its implications [81].

Based on this review, no model has been identified as being consistently adequate. To compare the performance of all models identified in our review, preferably we require IPD from multiple studies and in each study all the predictors within all the different models are recorded. Then it would be possible to compare the models directly across multiple settings. Only the paper by Amir et al. [56] tried to perform such a comparison (Table 2.3), which assessed the performance of models on the same data set from South Manchester, UK. The population of 4536 women had been assessed in a hospital clinic for breast and other cancer risks and were a high risk sample. Amir et al. [56] investigated the total and screened population; concluding that the Tyrer-Cuzick model was the most accurate for this high risk sample. It is not clear whether the Tyrer-Cuzick model would also be the best predictive model for a general population sample (a general setting), but this study is helpful as it gives an insight as to how the models perform on the same dataset.

2.6.5 Difficulties of doing an aggregate data meta-analysis of risk prediction models

This review serves as an empirical study of the feasibility of doing a review and meta-analysis of risk prediction model studies. The findings observed are specific to the breast cancer literature, but are likely to generalise more widely. Clearly, meta-analysis based on the published evidence is difficult as reporting standards are poor, validation statistics and parameter estimates are often not available, and most models do not receive adequate external validation. Thus key information desired, such as C statistics, E/O statistics, parameter estimates, standard errors and CIs are often not available, and so meta-analysis is limited. Furthermore, different models are proposed by different studies, and these are then validated in different studies, with many not receiving any validation. Furthermore, many models are developed using inferior methodology. For example, continuous variables often dichotomised in the model and non-linear trends are not considered [15]. Thus, being reliant on the standards of analysis and reporting of the primary studies is a major problem for synthesis of existing risk prediction studies.

2.6.6 What would be the advantages of IPD?

Many of the aforementioned issues would be overcome if IPD were available. For example, if an article were to present just the results and no confidence intervals or standard errors, then it would still be possible to use the IPD from this study to derive the missing information. This would allow the computation of the E/O ratio and C statistic for every article if IPD were available. The causes for high levels of between study heterogeneity could be explored if IPD were available. Even if an article did not present any validation results, it would be possible to use the IPD and calculate the

validation results and their uncertainty. Sometimes one may be able to validate additional models, in addition to those that were originally considered in the validation paper, if the relevant predictors are recorded in the IPD. Thus, if IPD were available it would make it much easier to reach an overall conclusion or get closer to which model performs the best. With IPD one could also consider re-estimating a model using all the available studies, to pool parameter estimates from all studies in meta-analysis and investigate heterogeneity more easily.

2.6.7 Next steps for the thesis

Aside from the findings about the breast cancer risk prediction models themselves, this review has identified interesting methodological issues for meta-analysis of risk prediction models. For example the poor quality of reporting and the often lack of validation statistics (or presenting them without CI or SEs), limits meta-analysis considerably. Therefore evidence synthesis of risk prediction models is important but potentially problematic using the traditional systematic review approach and when attempting to do a meta-analysis based on extracted results. Ideally evidence synthesis would be improved by using IPD from multiple studies, so not to be reliant on extracting results and to enable multiple models to be evaluated in the same datasets. To illustrate this, in Chapter 4 statistical methods for an IPD meta-analysis will be evaluated, in the context of developing and validating a model for predicting who develops hypocalcaemia in patients with a thyroidectomy. Before this, in Chapter 3 a review will be performed on how others have developed and/or validated a prediction model using IPD from multiple studies, so to identify current practice and ascertain whether the IPD lives up to the promise.

2.6.8 What this chapter adds?

Table 2.5 summarises the key findings of this chapter, and the review has been published in the journal Breast Cancer Research and Treatment [41].

Table 2.5: What chapter two adds

What is known / what is the problem?
<ul style="list-style-type: none">• Many breast cancer risk prediction models published, but rarely compared against each other• How do we determine which is the best model to use?• How do we compare the performance of these models?• These models are poorly reported (or not at all) and validation statistics are not always reported
What this study adds?
<ul style="list-style-type: none">• The models themselves were not available for comparison• This research has used a novel approach of comparing models by meta-analysing discrimination and calibration statistics, and presenting their average performance across different populations• This research indicates proposed models calibrate well on average and the predictors used in these models appear important, so they should help to inform clinicians and patients how they can reduce their breast cancer risk• However, discrimination is rather weak and further research may look at the inclusion of additional predictors• There is a need to validate these models in either similar or the same population to then be able to judge which of the models are better in comparison
What is needed next?
<ul style="list-style-type: none">• The papers that publish these models and their results need to improve the quality of their reporting, i.e. the models should be presented in their entirety• For those models that have been validated (internally and ideally externally), the validation statistics should be presented for discrimination and calibration• More validation studies are required for these models to compare directly in the same populations• If raw data and the models were available, then it would be possible to calculate model estimates and validation statistics, raw data and models need to be made available

CHAPTER 3: SYSTEMATIC REVIEW OF ARTICLES DEVELOPING OR VALIDATING A PREDICTION MODEL USING IPD FROM MULTIPLE STUDIES

3.1 Introduction

The articles on risk prediction models usually report model development, but only a few report external validation, and this was evident in the breast cancer review of Chapter 2. Lack of external validation could be a possible explanation as to why few models are being adopted in practice when so many are being developed [19]. The approach of using IPD from multiple studies offers a novel opportunity to overcome this lack of validation, for example using some of the studies to develop the model and use the remaining studies to validate the model.

There has been relatively little methodological research in relation to using IPD from multiple studies for development and validation of risk prediction models, though interest is growing [19, 82]. Thus it is important to ascertain what researchers are doing in current practice when IPD from multiple studies are available. For example, do they treat IPD as if all coming from one study? How do they identify which IPD to use? Do they assess heterogeneity in predictive factor effects? Do they keep back some studies for validation, and if so, how many? Is model performance checked in each study separately, or just one study?

The aim of this chapter is to perform a qualitative review to assess how risk prediction models are being developed and validated when IPD are sought from multiple studies and then combined. The aim is to identify current research techniques and standards; the role of

IPD meta-analysis methods towards development and validation; and the methodological challenges and problems faced by researchers. This review will allow recommendations for how research can be improved and flag those methodological techniques and issues researchers should recognise when modelling risk prediction using IPD. The findings will also direct research in the remainder of the thesis.

This review has now been published in *BMC Medical Research Methodology* (Ahmed et al. [16]) and I am the first author as this work is based entirely on the methods, results and recommendations of this chapter.

3.2 Methods

The review aims to identify and evaluate published articles that developed and/or validated a risk prediction model using IPD from multiple studies. The review methods are now described in detail.

3.2.1 Identifying potentially relevant articles

To identify potentially relevant articles, an existing database of 385 IPD meta-analyses articles was used. This database had already been formed by a systematic search in Medline, Embase and the Cochrane library using a search strategy previously described [83, 84]. The articles in this database were published from 1991 to March 2009, and formed the largest collection of IPD meta-analyses already available when the review was under taken. The database contained IPD meta-analyses for any purpose (e.g. treatment effects, diagnostic tests), and thus contained those for risk prediction (see below).

Note that the aim was not to review an exhaustive set of all risk prediction research using IPD from multiple studies, but rather to identify the main methods used and the key

methodological limitations and challenges. It was considered that qualitative saturation could be achieved with the existing database, and therefore it was not considered necessary to update the review with articles published since 2009.

3.2.2 Inclusion and exclusion criteria

A relevant article was defined as one which developed and/or validated a risk prediction model (i.e. a model aiming to predict individual outcome risk using single or multiple prognostic factors) using IPD from multiple studies. Note that the review does not consider articles which used an existing database containing multiple sources (e.g. practices). This has similar issues, but the focus here is on situations where IPD was firstly obtained from multiple studies and then meta-analysed (the standard framework for an IPD meta-analysis).

The term ‘model’ is used here for any developed equation, tool, or classification approach that allowed an individual’s risk to be predicted. There were no restrictions on the type of outcome being predicted or baseline disease/health of the patients under investigation, or the types of study (observational studies, randomised trials etc.) being utilised. Articles that aimed to identify factors which predict treatment effect were excluded, as treatment effects were of no interest. Articles that evaluated one or more factors for prognostic ability, but not in relation to absolute outcome risk were excluded (i.e. prognostic factor studies were excluded) [84].

Two independent reviewers (Ikhlaaq Ahmed and Dr Thomas Debray) screened and classified the abstracts and titles of each of the 385 articles. The articles were classified in regards to the risk prediction model status as either ‘yes’, ‘unsure’, or ‘no’. Any discrepancies that arose between the two reviewers were identified by me, and then solved through discussion if possible. A third reviewer, Dr Richard Riley then checked the ‘yes’ and any remaining

‘unsure’ articles, and a random 10% sample of the ‘no’ articles. Any further discrepancies between the three reviewers were resolved through discussion after reading the full papers and a final decision was then made.

3.2.3 Data extraction and in-depth evaluation of articles

Each article classified as a ‘yes’ was used for in-depth evaluation. A data extraction form was developed that included 77 questions. These 77 questions covered the rationale, conduct, analysis, reporting, and feasibility of the project developing and/or validating a risk prediction model using IPD from multiple studies. These questions formed the data extraction plan of the review (see APPENDIX A for a complete list of questions), which was documented and agreed upon by all authors in advance of the review.

I read each article in full and extracted information that answered these 77 questions, and then Dr Debray also independently answered these questions for each article. Both sets of extractions were compared and any discrepancies were identified by me, and resolved through discussion with Dr Riley and Dr Debray.

A summary of the questions used to evaluate each article is as follows:

Background

- Questions to ascertain the background information presented in the articles that relate to the central location of the IPD projects, the protocol, and ethics approval.

Objectives

- Questions to establish the research objectives, the baseline condition of the patients and the outcome being predicted.

Identifying IPD studies

- Questions to determine how they identified the relevant studies for inclusion, what types of studies were included in the meta-analysis, (RCT, observational etc.) and if they determined the number of studies required or the sample size for model development and/or validation.

Obtaining IPD

- Questions to establish how they asked for and obtained IPD, how the relevant study authors were approached, how many were approached and what proportion of these actually provided IPD, and whether there was any assessment regarding study quality
- Questions to assess if some of the requested IPD was not obtained and if any reasons were stated for this, and whether the number of patients/events was described within each study (if not, then overall).

Missing data

- Questions to ascertain whether missing data was considered; for example, if data was missing for patients and/or studies, and how any missing data was handled.

Model development

- Questions to determine how they developed their models. In particular, whether a statistical plan was presented for this process; how they handled data coming from multiple studies; what statistical models were used, and whether heterogeneity

between studies was considered. Also, how continuous predictors were handled, what criteria and procedure were used for inclusion of a predictor in the model, and whether the final model (e.g. with alphas, betas and their uncertainty) was presented.

Model validation

- Questions to establish how they validated (if at all) their prediction model, both internally and externally; the statistics used and presented, and the number of patients and number of studies used towards the validation; and whether bootstrapping or shrinkage was considered to assess model accuracy or adjust for over-optimism.

Bias

- Questions to ascertain if they stated anything in regards to those studies not willing/able to provide IPD; in particular whether there were any qualitative/quantitative differences between those studies that provided IPD and those that did not; whether the number of patients or events were stated from those studies not contributing to the analysis; and any suggestion that the studies obtained were a biased set of studies (if appropriate).

Conclusions

- Questions to determine the main conclusion from the analysis, and any limitations and problems discussed.

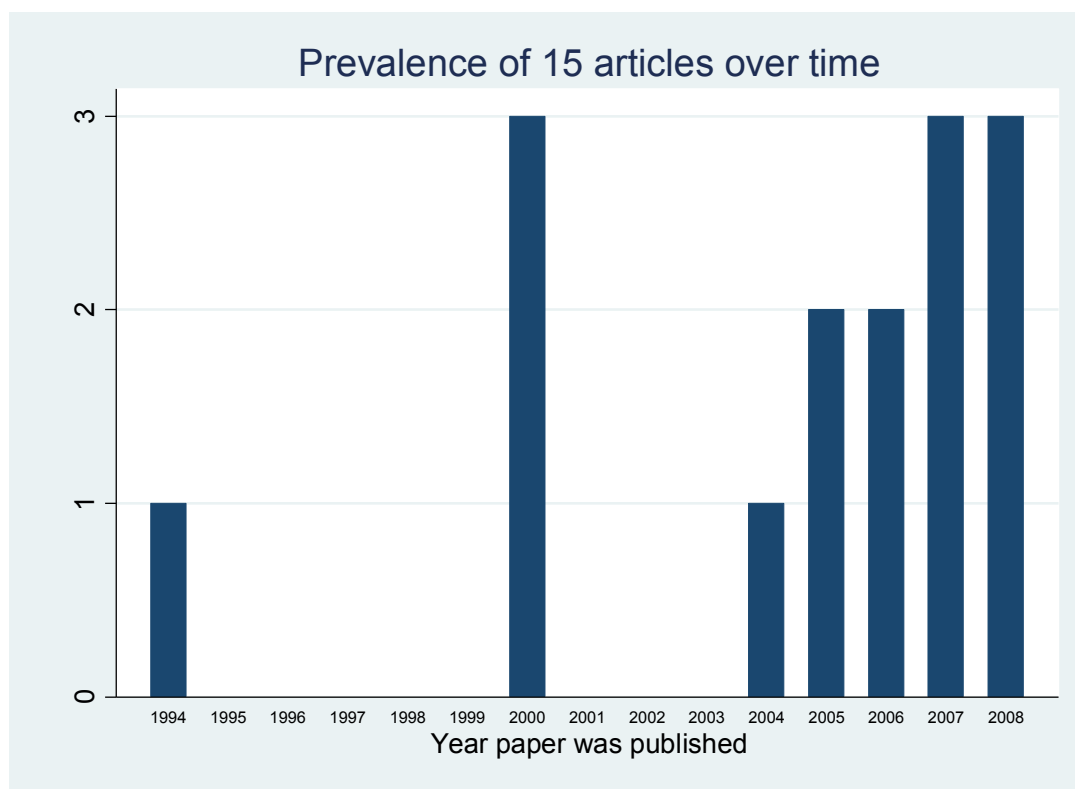
3.3 Results

The key findings of the review are now summarised.

3.3.1 Classification results

The classification process of the 385 articles identified 15 relevant articles that each developed and/or validated a prediction model using IPD from multiple studies [85-99]. They were published between 1994 and 2008. All 15 of these articles developed a model, and 11 also undertook some form of validation of their model [85-88, 90, 92-95, 98, 99]. See Figure 3.1 for publication dates for these 15 articles.

Figure 3.1: Graph showing distribution of the year of publication of the 15 articles identified



3.3.2 Background information

The central location of the 15 IPD projects (where the first author was located) included North America and Australasia, but most of the articles were from Europe (11); mainly Netherlands (4) and the UK (3). Just three of the 15 articles referred to a protocol for their IPD project (e.g. Yap et al. [99]), and only six of the articles mentioned ethics approval (e.g. Heffner et al. [86]), see Table 3.1 for full details.

3.3.3 Research objectives

All 15 articles stated their key research aims in relation to risk prediction; in particular the condition of the patients assessed at baseline and the clinical outcomes or onset of diseases of interest for prediction (see Table 3.1). Thirteen articles included patients who were diseased at baseline (e.g. Sylvester et al. [93], who predicted the probability of recurrence and progression at one and five years in patients with superficial bladder cancer) and two articles included patients who were healthy at baseline (e.g. Fowkes et al. [97], with outcome of interest being total and cardiovascular mortality in healthy individuals). The diseases considered at baseline included pancreatic cancer, chronic hepatitis C, bladder cancer and post-myocardial infarction among others (see Table 3.1). The outcomes predicted were either general (e.g. mortality) or disease-specific (e.g. development of radiation myelopathy, fatal coronary heart disease (CHD) within 10 years and postoperative symptomatic hypocalcaemia).

Table 3.1: Background information for each article

Paper	Author location	What was the key research aims of the paper in relation to prognosis/risk prediction?	At baseline what was the condition of the patients being assessed?	What outcomes or diseases were of interest for prediction?	Approach used to identify studies?	Number of studies providing IPD / Number of studies approached	Was the number of patients within each of the IPD studies given?
1994 Pagliaro[85]	Italy	To identify predictors of short-term and sustained Alanine transaminase(ALT) normalization after interferon treatment in adult patients with hepatitis C	Adult patients with transfusion-related chronic hepatitis C (trial 1) Adult patients with community-acquired chronic hepatitis C (trial 2)	Short term and sustained response (ALT normalization)	Collaborative group	2	Yes
2000 Heffner[86]	USA	To determine the predictive accuracy of pH for identifying patients with malignant pleural effusions who will fail pleurodesis	Patients with malignant pleural effusions	Failure of pleurodesis	Literature review	6/12	Yes
2000 Raboud[87]	Canada	To determine the ability of intermediate plasma viral load (pVL) measurements to predict virologic outcome at 52 weeks of follow-up in clinical trials of antiretroviral therapy	Patients with CD4 cell counts between 200 and 600 cells/mm3, naive to antiretroviral therapy and not been previously diagnosed with AIDS (INCAS) Patients had CD4 counts between 150 and 300 cells/mm3, naive to antiretroviral therapy, including nucleoside and nonnucleoside analogues or protease inhibitors.(AVANTI-2, AVANTI-3)	Virologic outcome at 52 weeks of follow-up	Collaborative group	3	Yes
2000 Terwee[88]	Holland	To develop a prognostic tool for patients with unresectable pancreatic cancer to distinguish between with low or high probabilities of survival 3 to 9 months after diagnosis.	Patients diagnosed with pancreatic cancer	Overall survival	Literature review	8 /15	Yes
2004 Chau[89]	United Kingdom	To identify baseline patient- or tumour-related prognostic factors To assess whether pre-treatment quality of life (QoL) predicts survival in patients with locally advanced or metastatic esophago-gastric (EG) cancer.	Patients with histologically confirmed inoperable adenocarcinoma, squamous cell carcinoma (SCC), or undifferentiated carcinoma of the oesophagus, esophago-gastric junction(EGJ), or stomach; adequate hematologic, renal, and hepatic function and Eastern Cooperative Oncology Group performance status (PS) 0 to 2.	Overall survival	Collaborative group	3	Yes
2005 Horn[90]	Holland	Investigate whether Transcranial Doppler (TCD) monitoring for micro embolic signals (MES), directly after carotid endarterectomy (CEA) may identify patients at risk of developing ischaemic complications.	Carotid endarterectomy patients (CEA)	Cerebral ischaemic complications, defined as new neurological deficits(amaurosis fugax, transient ischemic attack(TIA), minor and major ischaemic stroke) developing within the 1st week after CEA	Literature review	7/10	Yes
2005 Nieder[91]	Germany	Identifying the predictive value of biologically effective dose as function of the risk of myelopathy	Patients with spinal cord retreatment	Development of radiation myelopathy	Literature review	8/8	Yes
2006 Asia Pacific[92]	Australia	To investigate the generalisability of current definitions of the metabolic syndrome in Asia-Pacific populations, and	Healthy patients aged 30–75	Fatal CHD within 10 years	Collaborative group	26	Yes

Paper	Author location	What was the key research aims of the paper in relation to prognosis/risk prediction?	At baseline what was the condition of the patients being assessed?	What outcomes or diseases were of interest for prediction?	Approach used to identify studies?	Number of studies providing IPD / Number of studies approached	Was the number of patients within each of the IPD studies given?
2006 Sylvester[93]	Belgium	determine the prognostic value of metabolic risk factors to discriminate fatal coronary heart disease (CHD) risk To predict a superficial bladder cancer patient's probability of recurrence and progression at one and five years	Stage Ta, T1, and Tis bladder cancer patients who have undergone transurethral resection (TUR)	Time to first recurrence (disease-free interval) and time to progression to muscle invasive disease	Collaborative group	7	Yes
2007 Noordzij[94]	USA	Early prediction of hypocalcaemia after thyroidectomy using parathyroid hormone	Patients undergoing thyroidectomy	Postoperative symptomatic hypocalcaemia	Literature review	9/15	Yes
2007 Rovers[95]	Holland	To determine the predictors of a prolonged course for children with acute otitis media (AOM) to discriminate between children with and without poor outcomes	Children with acute otitis media	The primary outcome was a prolonged course of AOM, (defined as pain and/or fever at 3 to 7 days)	Literature review	6 /10	No, only overall given.
2007 Schaich[96]	Germany	To identify prognostic indicators in acute myeloid leukaemia (AML) to provide a new prognostic model for risk stratification of AML patients with +8	AML patients	Overall survival and relapse-free survival	Collaborative group	8	Yes
2008 Fowkes[97]	United Kingdom	To determine if the Ankle Brachial Index (ABI) provides information on the risk of cardiovascular events and mortality independently of the Framingham Risk Score (FRS) and can improve risk prediction	Participants of any age and sex derived from a general population	Total and cardiovascular mortality	Literature review	16/20	Yes
2008 Steyerberg[98]	Holland	To develop prediction model for predicting unfavourable outcome according to the Glasgow Outcome Scale (GOS) at 6 months after injury	Patients with moderate and severe TBI (GOS \geq 12)	6-months mortality and unfavourable outcomes defined by 6 months GOS	Collaborative group	11	Yes
2008 Yap[99]	United Kingdom	To design a prognostic indicator using demographic information to select patients at risk of dying after Myocardial Infarction (MI)	Patients at day 45 post-MI up to 2 years	All-cause, arrhythmic and non-arrhythmic cardiac mortality within 2 years	Not stated	4 /not stated	Yes

3.3.4 Identifying studies

Fourteen of the 15 articles stated the process they used to identify the studies for the IPD project (see Table 3.1), either a literature review or a collaborative group approach. A literature review is a review process where studies are identified through a keyword search on one or several databases containing potentially relevant literature. A collaborative group approach is a non-systematic approach where the researcher is part of a group, or has a contact within a group, and this group has access to IPD from multiple studies that are relevant for inclusion. Seven articles used a collaborative group approach, e.g. Chau et al. [89], Asia Pacific Cohort Studies Collaboration [92] and Pagliaro et al. [85].

Seven articles used a literature review approach, and all of these described the databases and sources searched (e.g. Terwee et al. [88]). However, only Fowkes et al. [97] and Noordzij et al. [94] gave the keywords used within this search, and only Fowkes et al. [97] gave a flowchart of the process of searching, classifying and retrieving IPD studies. Only three of the seven articles described how authors of identified studies were approached for their IPD and this included e-mail, postal mail, and telephone.

All 15 articles did not calculate a required sample size for model development and/or validation; i.e. they just sought all IPD available from the studies identified in the review or the studies available in the collaboration.

3.3.5 Obtaining IPD from multiple studies

All seven literature review articles gave the number of studies they obtained IPD from and the number of studies they approached for IPD. Six of the seven articles did not obtain IPD for all studies desired; this raises concern of availability bias. Availability bias relates to

when studies that provide their IPD are a biased subset of all existing studies [100]. The percentage of IPD obtained ranged from 50% to 100% (see Figure 3.2). Only three articles gave reasons as to why they could not obtain IPD for the desired studies. For example Heffner et al. [86] stated that data was no longer available or not saved for those studies not providing IPD.

Figure 3.2: IPD obtained and total IPD studies desired in the seven articles using a literature review to identify relevant studies

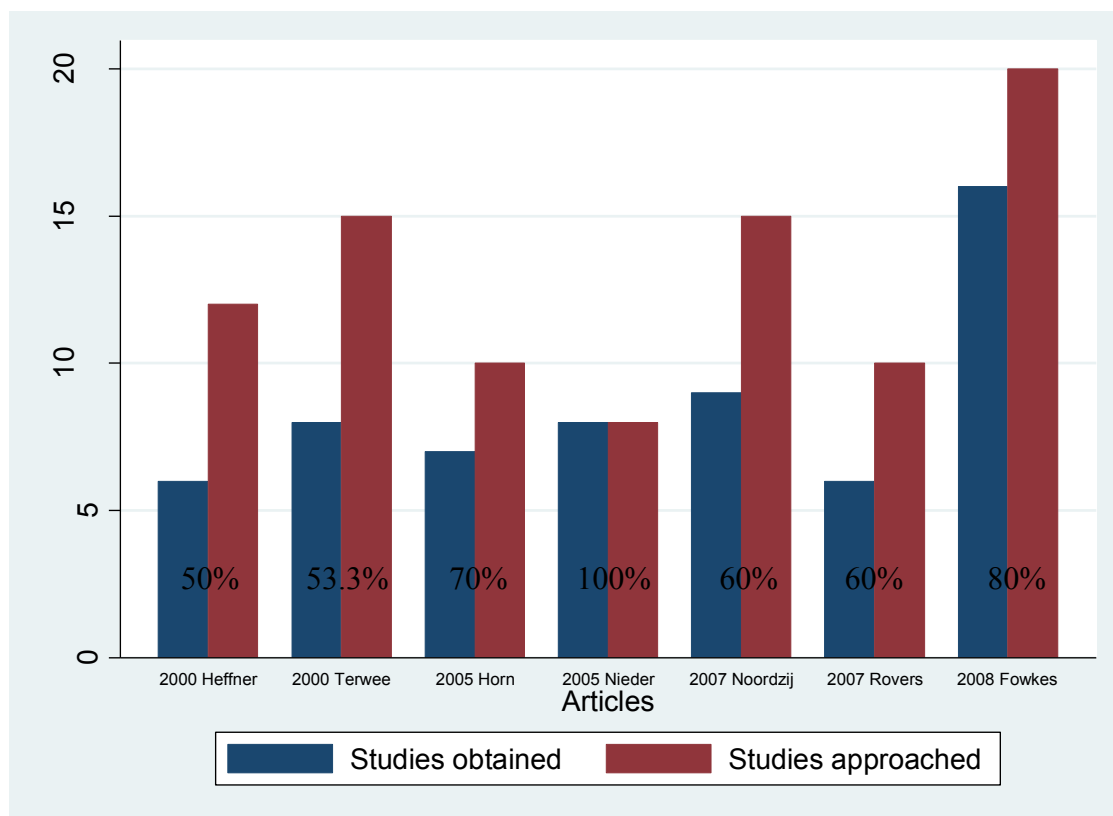


Table 3.2: Summarising the IPD

Paper	Types of studies included:	Was the number of patients within each of the IPD studies given? If not, was it given overall?	Number of events per study given? If not, was it given overall?	Was the number of events per candidate variable (predictor) given in each study? If not, was it given overall?
1994 Pagliaro[85]	RCT	Yes	Yes	Yes
2000 Heffner[86]	Observational	Yes	Yes	No
2000 Raboud[87]	RCT	Yes	No	Yes, there is only 1 predictor
2000 Terwee[88]	Follow-up studies (unclear)	Yes	Yes	Yes, table 2 for overall
2004 Chau[89]	RCT	Yes	No, only overall given	Yes
2005 Horn[90]	Not clear	Yes	Yes	No, only overall given
2005 Nieder[91]	Not clear	Yes	Yes	Yes (the actual IPD is given)
2006 Asia Pacific[92]	Observational	Yes	No, only overall given	Yes
2006 Sylvester[93]	RCT	Yes	No, only overall given	No, only overall given
2007 Noordzij[94]	Observational	Yes	No, only overall given	Yes
2007 Rovers[95]	RCT	No, only overall given.	No, only overall given	Yes
2007 Schaich[96]	RCT	Yes	No, only overall given	Yes
2008 Fowkes[97]	Observational	Yes	Yes	Yes
2008 Steyerberg[98]	RCT (8) and observational (3)	Yes	No, only overall given	No, only overall given
2008 Yap[99]	RCT	Yes	Yes	No

3.3.6 Details of IPD obtained

The types of studies providing IPD were stated by 12 of the 15 articles (Table 3.2). Seven articles used IPD from RCT's (e.g. Pagliaro [85]), four used observational studies (e.g. Heffner et al. [86]), and Steyerberg et al. [98] used both RCT's and observational studies. In the eight articles that included RCTs, five used data from all treatment groups, two only used the placebo group, and Steyerberg et al. [98] used all the data in their primary analysis but just the placebo group for their secondary analysis.

A summary of the sample population (i.e. baseline patient characteristics) was given *separately* for each IPD study in eight of the 15 articles, (e.g. Pagliaro [85]), whilst five articles just gave a summary across the overall, combined IPD (e.g. Sylvester et al. [93]); two articles did not give this information at all (e.g. Raboud et al. [87]). Only one of the 15 articles (Rovers et al. [95]) mentioned assessing the quality of studies; though it was not clear if this had any implications for the IPD they analysed.

Fourteen of the 15 articles gave the number of patients within each of the IPD studies, and Rovers et al. [95] gave just the overall number of patients across all studies (Table 3.2). Seven of the 15 articles gave the total number of events for each predicted outcome within each of the IPD studies (e.g. Yap et al. [99]); whilst seven articles just gave the overall number of events across all studies (e.g. Noordzij et al. [94]) and one article did not give this information at all (Raboud et al. [87]). Nine of the 15 articles reported the number of events per candidate predictor in each IPD study (e.g. Pagliaro [85]), whilst three articles gave this information for the overall IPD across all studies (e.g. Terwee et al. [88]), and three articles did not report this information.

3.3.7 Missing data

Eleven of the 15 articles mentioned missing data (see Table 3.3). Eight of the 15 articles mentioned having missing patient data for some predictors; for example in Noordzij et al. [94] only six of the nine studies that supplied IPD obtained preoperative PTH values, whilst Steyerberg et al. [98] state “pupillary reactivity was not recorded in two trials”.

If an article reported that missing patient data occurred for a variable (candidate predictor), then generally the article either excluded patients with missing data from the analysis (e.g. Heffner et al. [86]), or used a (multiple) imputation approach (e.g. Rovers et al. [95]). Additionally three of the 15 articles entirely deleted some studies due to absence of a predictor or outcome of interest; for example Terwee et al. [88] excluded three of the eight studies, and state: “one study was excluded because we could not distinguish between patients with tumours of the pancreatic head and tumours of the pancreatic body or tail, and two others were excluded because information on metastases was lacking or incomplete”.

Table 3.3: Details of missing data in the IPD projects and how it was handled

Paper	Was missing data mentioned/indicated within the article?	If stated, how was missing data handled in the analysis?
1994 Pagliaro[85]	Not stated	NA
2000 Heffner[86]	Yes	Missing values were excluded from the analysis. Analysis was performed and then repeated with an undefined imputation approach (similar results)
2000 Raboud[87]	Yes	Those patients were excluded from ROC curve (analysis)
2000 Terwee[88]	Yes	Missing values for age were imputed by using the sample mean
2004 Chau[89]	Yes	Those patients were excluded from the prognostic model
2005 Horn[90]	Not stated	NA
2005 Nieder[91]	Yes	Those patients were excluded from the analysis
2006 Asia Pacific[92]	Yes	Those studies that did not measure each of those risk factors at baseline were excluded from the analysis
2006 Sylvester[93]	Yes	Not stated
2007 Noordzij[94]	Yes	Those three studies were excluded from the analysis
2007 Rovers[95]	Yes	Missing data was imputed for each trial by using the linear regression method (missing value analysis)
2007 Schaich[96]	Not stated	NA
2008 Fowkes[97]	Yes	Missing data was imputed using the expectation-maximization procedure for multivariate normal data
2008 Steyerberg[98]	Yes	Missing data was imputed using Multivariate Imputation by Chained Equations (MICE) algorithm
2008 Yap[99]	Not stated	NA

3.3.8 Model development

The statistical analysis methods used for model development are now summarised. The main focus here is how IPD from multiple studies were handled.

3.3.8.1 Analysis method

The number of patients and events used from each study separately towards the prediction model development was stated for 10 of the 15 articles (e.g. Yap et al. [99]), with the remaining articles focussing on the overall numbers.

A statistical analysis plan for prediction model development was stated in the methods section for all 15 articles (see Table 3.4). The two articles that considered a single predictor in their model examined predictive performance by calculating ROC curves from the data. In the 13 articles developing a multivariable model, six used a Cox regression model (e.g. Terwee et al. [88]), five used a logistic regression model (e.g. Heffner et al. [86]), and the Asia Pacific Cohort Studies Collaboration [92] used both Cox and logistic regression models.

All 15 articles used IPD from multiple studies to develop their models. Of these 15 articles, 10 developed the risk prediction model via ‘one-step meta-analysis ignoring clustering’ which is an unstratified approach, where authors (e.g. Horn et al. [90]) pool all the data together into one big dataset, and ignore clustering by study or collaborative group, usually *without* explaining why clustering was ignored. A notable exception is Terwee et al. [88], who justify their approach by examining whether stratification by study was necessary, stating: “The homogeneity assumption was checked by including treatment as a dummy-coded variable into both models (stratification per study). Initially we found a significant survival benefit for patients treated by surgical bypass procedures compared with endoscopic stents. However, this effect disappeared after adjustment for Karnofsky’s index (a measure of functional status) in the study in which this variable was available, which legitimises pooling.”

Three of the remaining five articles (e.g. Steyerberg et al. [98]) developed their model via the ‘one-step meta-analysis accounting for clustering’, where the data from all studies/collaborative groups are analysed together but with clustering by study/group accounted for (e.g. using a dummy variable for study). In another article, Fowkes et al. [97] developed their model using ‘a two-step approach’, where the data are first analysed separately in each study, and then their model estimates are pooled together in the second-step. They looked at mortality by study and use a two stage approach to look at the predicted value of ABI (Ankle Brachial Index) in addition to FRS (Framingham Risk Score). They estimated Kaplan-Meier curves for each combination of ABI (four categories) and FRS (five categories) in each study, and then performed a random effects meta-analysis across studies of the survival percentages obtained at different time-points (to give an average predicted survival percentage). In the remaining article, Schaich et al. [96] used a hierarchical cluster

approach, but it is not clear whether this cluster approach specifically included clustering by study.

None of the 15 articles developed their model using only a portion of the IPD available, i.e. no IPD was kept separate for validation, and so it was always all used for development.

3.3.8.2 Heterogeneity of predictor effects

Only three of the 12 articles considered between-study heterogeneity in the predictors within the prediction model (Table 3.4). Of these, Steyerberg et al. [98] state: “Similarly, study-specific effects were assessed with interaction terms between study and each predictor.

Interaction terms between predictors were examined with likelihood ratio tests, but none was of sufficient relevance to extend the models beyond the main effects for each predictor”.

Rovers et al. [95] assessed heterogeneity through the I^2 statistic and state: “To determine whether pooling was justified, heterogeneity between studies was assessed with the I^2 statistic. Because the I^2 value was 25%, pooling was performed”. Fowkes et al. [97] performed a test of heterogeneity in Ankle Brachial Index (ABI) (using Chi-square test and I^2), for one of their predictors; they then applied random effects in their analysis, which was likely due to heterogeneity being detected, although not explicitly stated.

3.3.8.3 Handling of continuous predictors

Continuous predictors were analysed on a continuous scale in four of the 15 articles. For example, Steyerberg et al. [98] state: “For the continuous predictor’s age, glucose, and Hb, a linear relationship with outcome was found to be a good approximation after assessment of non-linearity using restricted cubic splines”. The remaining 11 articles either categorised or dichotomised the continuous predictors of interest. For example, Heffner et al. [86] state that “continuous variables were entered as dichotomous indicator variables with test thresholds

determined by ROC analysis”, and Chau et al. [89] state “Laboratory variables were initially coded as continuous variables and subsequently dichotomised with the cut-off points chosen at the median value of each variable”.

3.3.8.4 Need for standardisation

Three of the 15 articles mentioned the need to standardise the coding of predictors and outcome definitions. For example, Terwee et al. [88] state “To standardise definitions among the different studies, the presence of metastases and of pain and weight loss at diagnosis were classified as ‘present’ or ‘absent’”, whilst Noordzij et al. [94] handled different methods of measuring PTH values by analysing percentage change in PTH from baseline, rather than analysing PTH on its original scale. In order to use and compare candidate predictor factors measured on different continuous scales, Steyerberg et al. [98] standardised the reporting of odds ratios so that they corresponded to a change from the 25th percentile to the 75th percentile of the predictor distribution.

3.3.8.5 Strategy for inclusion of predictors

Of the 13 articles that developed a multivariable model (i.e. a model with two or more included predictors), the most common approach (used by 6 articles) was to use p-values to decide which predictors were included (Table 3.4). For example, Heffner et al. [86] state: “Variables that were found by univariate analysis to be associated with pleurodesis failure with a value for $p < 0.10$ were entered into a logistic regression model. Continuous variables were entered as dichotomous indicator variables with test thresholds determined by ROC analysis. Variables were removed from the model if their p values were ≥ 0.05 ”. Five other articles used a selection procedure (Table 3.4), for example Pagliaro [85] used a stepwise procedure, and Chau et al. [89] used forwards and backwards selection procedures.

3.3.8.6 Reporting of developed model

In the 13 articles that developed a multivariable model, the model was adequately reported in terms of the predictor effects (e.g. with the six articles using the Cox model reporting either log hazard ratios or hazard ratios). However, none of the six articles that used logistic regression reported the alpha term (baseline risk) for their final model, with all six presenting the odds ratio (either as log odds ratio or on its original scale). However, Steyerberg et al. [98] do present a simplified version of their logistic regression model (based on a simple score chart), in which they do present both the alpha and beta terms.

Table 3.4: Model development

Paper	Statistical analysis plan for model development given?	In developing the prediction model, how was IPD from multiple studies synthesised?	Details of the approach (or statistical models) used in synthesising IPD from multiple studies:	Was between-study heterogeneity in the predictors/alpha term considered within the prediction model? If so how (e.g. using random-effects in the analysis; assessing heterogeneity using I-squared)	What criteria were used to decide inclusion of a predictor in the model? (e.g. statistical criteria, such as $p < 0.1$, or clinical criteria such as a hazard ratio > 2 or inclusion of 'smoking' variable regardless)	What selection procedure was used?	How was the final model reported?
1994 Pagliaro[85]	Yes	One-step ignoring clustering	Logistic regression	No	Significance of each predictor was checked by maximum likelihood approach with 2 tailed p-values	Stepwise	The beta, SE(beta), 95%CI and p-values for each variable were reported
2000 Heffner[86]	Yes	One-step ignoring clustering	Logistic regression	No	$P < 0.1$ was used to find significance in univariate analysis, and those that were significant were included in the logistic regression model. Variables were removed from the model if their p values were > 0.05	P values	The final model consists of only 1 variable. The OR and 95% CI are stated for its predictor
2000 Raboud[87]	Yes	One-step ignoring clustering	ROC graphs	No	NA (only 1 predictor is considered)	NA	NA
2000 Terwee[88]	Yes	One-step ignoring clustering,	Cox regression	No	The variables age, sex, and presence of metastases were available for all studies and included in the model. Additionally, an extended model that also containing pain, jaundice, and weight loss at diagnosis was developed on a subset of the four studies that provided this information.	Not used	Relative risks with 95%CI are provided
2004 Chau[89]	Yes	One-step analysis accounting for clustering	Cox regression	No	Univariate assessment of the prognostic effect of each factor, leading to a multivariate analysis. A two-sided $P < 0.01$ was considered significant.	Forward and backwards stepwise	Hazard ratio with 99% CI, and risk ratios with 99%CI reported
2005 Horn[90]	Yes	One-step ignoring clustering	Logistic regression	No	Predictors with statistically significant univariate associations ($p < 0.05$) were included in the multivariate analysis	P values	OR's and 95% CI reported
2005 Nieder[91]	Yes	One-step ignoring clustering	Not stated	No	Not stated	Not stated	The weights of the risk score are presented alone
2006 Asia Pacific[92]	Yes	One-step ignoring clustering	Cox and logistic regression	No	All five "metabolic" risk factors at baseline were used	NA	Only variable names reported
2006 Sylvester[93]	Yes	One-step analysis accounting for	Cox regression	No	Variables that represented the prior recurrence rate, number of tumours,	P values	HR with CI reported

Paper	Statistical analysis plan for model development given?	In developing the prediction model, how was IPD from multiple studies synthesised?	Details of the approach (or statistical models) used in synthesising IPD from multiple studies:	Was between-study heterogeneity in the predictors/alpha term considered within the prediction model? If so how (e.g. using random-effects in the analysis; assessing heterogeneity using I-squared)	What criteria were used to decide inclusion of a predictor in the model? (e.g. statistical criteria, such as $p < 0.1$, or clinical criteria such as a hazard ratio > 2 or inclusion of 'smoking' variable regardless)	What selection procedure was used?	How was the final model reported?
		clustering			tumour size, T category, grade, and CIS were included in the final multivariate model. Age and gender were likewise not retained in the final models, as they were not significant at the 5% level ($P < 0.05$)		
2007 Noordzij[94]	Yes	One-step ignoring clustering	Sensitivity, specificity and ROC used	No	Only one predictor used, PTH	NA	NA
2007 Rovers[95]	Yes	One-step ignoring clustering	Logistic regression, fixed-effects	Heterogeneity was assessed (through I^2 statistic, but not mentioned on which variables). To justify whether pooling was appropriate, they assessed heterogeneity between studies, found the I^2 to be 25% and pooled the data	Predictors with $P \leq 0.10$ in the univariate analysis were included in the multivariate analysis. The model was reduced through exclusion of predictors with $P > 0.05$	Backwards selection	Odds ratios and CI of the predictors reported
2007 Schaich[96]	Yes.	Hierarchical cluster (unclear whether cluster was study)	Cox regression, fixed-effects	No	Not stated	Limited backwards selection was stated	Hazard ratios and CI only provided
2008 Fowkes[97]	Yes	Two-step approach	Cox regression, random-effects	Test on heterogeneity on ABI (using χ^2 and P) was performed and random effects were applied	NA	NA	Not reported
2008 Steyerberg[98]	Yes	One-step analysis accounting for clustering	Logistic regression	Yes, study-specific effects were assessed with interaction terms between study and each predictor. Interaction terms between predictors were examined with likelihood ratio tests, but none provided enough evidence to extend the models beyond the main effects for each predictor.	Combination of practical (widely available) and statistical (based on Nagelkerke) arguments to select variables. 3 different models are explored: a core model, an extended model (core + information on secondary insults) and a lab model (extended + additional information on glucose and haemoglobin)	NA (it approximates a forward approach, based on a restricted set of variables)	Predictors were reported with their odds ratios
2008 Yap[99]	Yes	One-step ignoring clustering	Cox regression	No, but they have justified pooling all the data due to the overall lack of significant interaction between risk factors and studies across the four populations	Only variables that were significantly associated with mortality were included	Not stated.	Only the variable names were reported

3.3.9 Model validation

Now considered is whether articles validated their developed model and, if so, how. Findings are summarised in Table 3.5. ‘Internal validation’ is referred to when the same data is used to validate the model as to develop it; and ‘external validation’ is referred to when different data is used to validate the model (e.g. data from other studies that were not used to develop the model). Another type of validation is ‘internal-external cross-validation’ (Royston et al. [101]), which is an external validation approach that allows validation studies to also be included in the model development in rotation. In short, this technique excludes one of the IPD studies from the available set, and the remainder are used to develop the prediction model; the excluded study is then used to validate the model externally. This process is repeated for each study omitted in turn, and this allows the consistency of the developed model and its performance to be examined on multiple occasions. If the model performs consistently well in all omitted datasets, then all IPD studies can be used for model development. However, if the model performs poorly in some datasets, this may indicate heterogeneous studies (populations) for which the model does not generalise to and may require exclusion.

3.3.9.1 Internal, external or internal-external validation

Four articles did not validate their model (Table 3.5), such as Chau et al. [89]. Nine of the other 11 articles used only an internal validation approach. For example, Noordzij et al. [94] calculated ROC curves to assess the discrimination accuracy of the single predictor using all the IPD available. However, a few of these nine also recognised that model accuracy may be over optimistic in the development data, and so tried to correct apparent performance (in the original data) with an optimism-adjusted performance estimate. For example, Pagliaro et al. [85] took a random test set of 100 patients from their development data of 261 patients; they

show that the area under the ROC curve was slightly lower in the test data than the full original data (e.g. model 1: AUC =0.728 (n=261) vs. AUC=0.659 (n=100)). Sylvester et al. [93] were the only authors to perform bootstrap re-sampling of all the IPD available and they then examined model performance in these samples, leading to a bias-corrected C statistic. None of the articles considered or mentioned shrinkage.

Two articles did use a form of external validation: Steyerberg et al. [98] and Yap et al. [99]. Steyerberg et al. [98] used both external validation and internal-external cross-validation. For external validation, they use IPD from a trial different to that used to develop the model, but note problems with missing variables: “We aimed to validate all models externally using data from selected patients in the CRASH trial. However, lab values were not recorded in this trial, nor were hypoxia, hypotension, and EDH. We therefore validated the core model, and a variant of the extended model, in which only the Marshall CT classification and presence of tSAH were added to the core model (i.e., the core + CT model).” Interestingly, their developed model was stratified by study and so had multiple intercepts to choose from. However, in their validation they appear to choose one intercept related to a single trial as “it represented typical proportions of mortality (278/1,118, 25%) and unfavourable outcome (456/1,118, 41%)”. When performing internal-external cross-validation, Steyerberg et al. [98] state: “AUC was calculated in a cross-validation procedure, where each study was omitted in turn. Results were pooled over the ten imputed datasets for eight studies with sufficient numbers for reliable validation ($n > 500$).” They found that cross-validation performance was better for the three observational studies but slightly lower for the RCTs.

Yap et al. [99] also used internal-external cross-validation, stating: “The use of several different studies provided an opportunity to validate the general method using data drawn

from a different patient population using an internal–external cross-validation system proposed by Royston et al. [101] that leave-one-out cross-validation on a cohort basis. We therefore, sequentially designated one of the trials as a test study. The remaining studies acted as a training data set, which was used to generate the risk scores as above.” They did not state their cross-validation results.

3.3.9.2 Reporting of validation performance criteria

The number of studies used for validation was reported in seven articles, and these articles also reported the number of events and patients from each study used towards the validation.

Validation performance criteria focussed on discrimination, calibration, reclassification or goodness of fit (Table 3.5). Of the 11 articles that examined validation, 10 gave discrimination statistics such as sensitivity, specificity, AUC and the C Statistic. Three articles reported calibration statistics such as Hosmer-Lemeshow goodness-of-fit).

Eight articles provided figures to show model discrimination, accuracy or calibration. The most common figure was an ROC curve (presented by six articles, e.g. Horn et al. [90], Figure 3.3). Steyerberg et al. [98] presented calibration figures depicting the agreement between predicted and actual outcome probabilities, and Yap et al. [99] presented survival curves for each test study, to illustrate the consistency of the model in identifying risk categories.

Figure 3.3: Horn et al. ROC curve

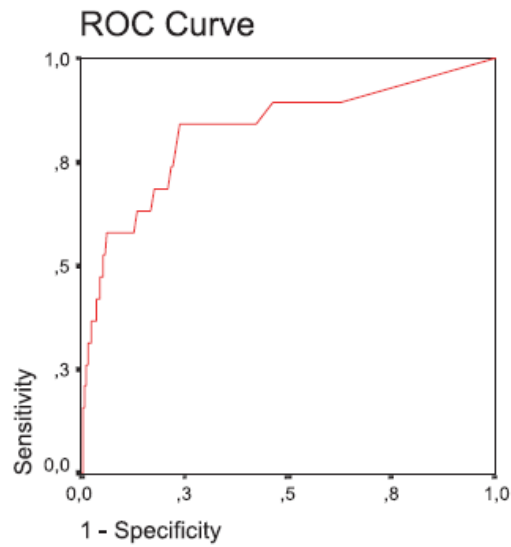


Figure 3.3 shows discriminative ability, as stated by Horn et al. [90]: *“To investigate the discriminative ability of MES monitoring and to determine the optimal cut-off value, a ROC curve was constructed, plotting sensitivity versus ‘one minus specificity’ for all possible cut-off values”*.

Table 3.5: Model validation methods and results reported in the articles

Paper	Validation type: (Internal or external)	How was this done?	How good was the performance of the model after the validation?	What discrimination, calibration, reclassification or other statistics (e.g. goodness of fit or R2) were used?	Were any figures given to show model discrimination / accuracy / calibration? If so, what?	Was the number of studies used in the validation stated?	Was the number of patients and events used from each study toward the prediction model validation given?
1994 Pagliaro[85]	Internal	Derived a random test set of 100 patients and calculated ROC curve and compared AUC to the remaining patients for both models.	AUC =0.728 (n=261) vs. AUC=0.659 (n=100) (model 1) and AUC = 0.703 (n=234) vs. AUC=0.673 (n=100) (model2)	Discrimination: AUC	Yes, Response rate charts and ROC curve	Yes	Yes, for each study
2000 Heffner[86]	Internal	Internal validation (limited descriptions)	Sensitivity& specificity are provided, although no AUC values given for the model.	Discrimination: AUC (for each predictor), Calibration: Hosmer-Lemeshow goodness-of-fit	Yes, a ROC curve is given for the predictor pleural fluid pH (the only predictor in the final model)	No	No
2000 Raboud[87]	Internal	The ROC curves were calculated for each predictor in the corresponding (and combined) datasets	The performance increased as the time approached the endpoint.	Discrimination: AUC, Sensitivity, Specificity	Yes, ROC graphs, and AUC trends for different follow-up times	Yes, the original studies are used (internal validation)	Yes
2000 Terwee[88]	Internal	ROC curve.	AUC = 0.75, 0.74, and 0.74 in the first 3, 6, and 9 months after diagnosis respectively	Discrimination: AUC	No	No	No
2004 Chau[89]	Not performed	NA	NA	NA	NA	NA	NA
2005 Horn[90]	Internal	ROC curve.	AUC = 0.83	Discrimination: AUC, Sensitivity, Specificity	Yes, ROC graph	No	No
2005 Nieder[91]	Not performed	NA	NA	NA	NA	NA	NA
2006 Asia Pacific[92]	Internal	AUC, ROC curves. The whole cohort of 5 studies was used	- AUC = 0.586 (modified NCEP-ATPIII definition) - AUC = 0.733 (model with 5 continuous risk factors) - AUC = 0.788 (model with age+sex)	Discrimination: AUC	Yes figure 2 shows ROC curve (graph)	Yes (although not explicitly)	Yes (although not explicitly)
2006 Sylvester[93]	Internal	To assess model accuracy (discrimination) at one and five years, C statistic was calculated using bootstrap re-sampling technique	C-stat = 0.74 at one year and 0.75 at five years.	Discrimination: C Statistic	No	Yes	Yes, same as derivation data
2007 Noordzij[94]	Internal	All data was directly used for validation	AUC = 0.967 (6 hours), AUC = 0.943 (2 hours)	Discrimination: AUC, Sensitivity, Specificity,	Yes, ROC curve	Yes	Yes

Paper	Validation type: (Internal or external)	How was this done?	How good was the performance of the model after the validation?	What discrimination, calibration, reclassification or other statistics (e.g. goodness of fit or R2) were used?	Were any figures given to show model discrimination / accuracy / calibration? If so, what?	Was the number of studies used in the validation stated?	Was the number of patients and events used from each study toward the prediction model validation given?
2007 Rovers[95]	Internal	ROC curves and goodness of fit Hosmer-Lemeshow test.	AUC = 0.63 for primary outcome. goodness-of-fit test with p-value of 0.93	PPV, NPV Discrimination: AUC, Calibration: goodness-of-fit	No	No	No
2007 Schaich[96]	Not performed	NA	NA	NA	NA	NA	NA
2008 Fowkes[97]	Not performed	NA	NA	NA	NA	NA	NA
2008 Steyerberg[98]	Internal-external cross-validation	Cross-validation procedure, where each study was omitted in turn and results were pooled over the ten imputed datasets for eight studies	AUC>0.80 in the 3 observational studies. Lower AUC in RCTs	Discrimination: AUC, Calibration: Hosmer-Lemeshow	Yes. There are figures depicting predicted and actual outcome probabilities	Yes, Internal-external(1 study at a time excluded), external was on CRASH RCT.	Yes
	External	Individual trial (CRASH) was used for external validation, although they develop the model stratifying by study, they take the intercept term from one trial and use that in their model to apply it; so they can validate it. Logistic regression was subsequently used to calibrate the risks of mortality and unfavourable outcome according to the scores, with the model intercept referring to the Tirilazad international trial	AUC = 0.776 and 0.780 for mortality and unfavourable outcome (core model). Lower AUC for the more advanced models. Calibration Hosmer-Lemeshow tests indicate miscalibration for core and core+CT models (p<0.001). Improved calibration for extended model p > 0.1 in particular group of patients				
2008 Yap[99]	Internal-external cross-validation	Internal-external cross-validation	There is no real measure of performance presented. There is a qualitative comparison of Kaplan-Meier curves	None given	Survival curves are presented for each split-validation sample, and illustrate consistency of the model in identifying risk categories	Yes	Yes

3.3.10 Dealing with those studies not willing / able to provide their IPD

In six of the seven literature review articles, IPD was not obtained for all desired studies (Table 3.1), but only three of these articles discussed this as a potential limitation of their project. For example Horn et al. [90] stated “We may not have identified all centres, which monitor MES after carotid surgery: it is likely that small series of patients remain unpublished and have therefore escaped our attention”. None of the six articles report the number of patients and events in the missing IPD studies, and only Rovers et al. [95] discussed the qualitative or quantitative differences between those studies providing IPD and those studies not able to. They were not able to include four trials, as IPD were not available for them; however, they found the excluded trials to be similar to those included in terms of aggregate data results, so they do not expect the results of the meta-analysis to change.

3.3.11 Conclusions and limitations

The developed risk prediction models had large potential for use in clinical practice, according to the discussion of the 15 articles. For example Noordzij et al. [94] state “PTH assay, when checked 1 to 6 hours after thyroidectomy, has excellent accuracy in determining which patients will become symptomatically hypocalcemic.” Similarly, Yap et al. [99] state "our study suggests that in post-MI patients, pre-selected using LVEF or frequent ventricular premature beats, the additional use of a simple prognostic indicator based on demographic and baseline information was able to segregate patients that were at high risk of dying, for 3 different modes of mortality."

There were also numerous limitations and problems of the IPD project noted in the articles' discussion sections. Some common limitations were: lack of standardisation in data collected between different studies, variability in definition of predictors, missing data and potential for

bias, lack of external validation, between-study heterogeneity, and loss of information due to categorisation or dichotomisation. The key limitations and methodological problems are summarised in Table 3.6.

Table 3.6: Limitations and methodological problems noted in discussion of the 15 articles

Paper	What limitations and problems of the IPD project were noted in the Discussion?
1994	- An untreated control group was excluded
Pagliaro[85]	- The predictor 'disease duration' is inaccurate
2000 Heffner[86]	- They constantly mention the limitations of the primary studies, their design and data - Plus potential for bias and missing data
2000 Raboud[87]	- Distortion of PPV and NPV due to no intention-to-treat - Selective drop-out
2000 Terwee[88]	- Pooling does not solve problems regarding the lack of standardisation in the measurement of risk factors as a result of retrospective designs, differences in diagnostic procedures, and missing data - Few variables available for all studies to enter in a prognostic model - Important risk factors may not have been included
2004 Chau[89]	- Prognostic index requires validation
2005 Horn[90]	- Small number of patients experiencing cerebral ischaemic complications shortly after surgery => large CI - Other baseline characteristics, which could possibly be risk factors were not analysed - The outcome measure was not ideal - Potential underestimation because assessment may not always have been performed by suitably qualified clinicians - Lack of external validation
2005 Nieder[91]	- Retrospective data analysis is subject to several sources of error and bias - Small number of patients, inconsistent use of chemotherapy, and uncertainty in dose calculation resulting from both the BED model and the differences in dose prescription and reporting among the different institutions in which the patients were treated - Potential bias in reported doses to the spinal cord in each individual paper - Differences in radiation sensitivity - Follow-up time may not have identified all patients with myelopathy - Validation should be undertaken
2006 Asia Pacific[92]	- Few cohorts measured all five risk factors at baseline, which limits the precision of our estimates and our ability to explore differences by age, gender and region - Most of these studies were initiated at time when the importance of waist circumference as a determinant of CHD risk was not well-recognized - Lack of standardization in data collection between studies, including in outcome ascertainment
2006 Sylvester[93]	- No external validation - Characteristics and prognoses of patients may have changed (old data: study published in 1996, data collected between 1979 and September 1989) - Recurrence and progression rates reported here may be higher than those found in current clinical practice
2007 Noordzij[94]	- Prospective outcomes studies need to be performed - Inherent variability that exists between various PTH assays and the variability in the definition of hypocalcaemia (outcome) used in the studies included in this meta-analysis - Considerable variability in the mean preoperative PTH values in the studies included in this analysis
2007 Rovers[95]	- Generalizability & power: 4/10 trials not included - Generalizability: only children from trials included and from observational arm - Bias: not all studies used objective diagnostic methods - Loss of information & bias due to dichotomization and completely missing variables
2007 Schaich[96]	- Small sample size - Time-selection bias - Heterogeneity
2008 Fowkes[97]	- No recalibration for FRS model
2008 Steyerberg[98]	- "Old" time period - Motor score is not always available and may be unreliable - Missing variables - Potential misclassification bias in dichotomizing the outcome
2008 Yap[99]	- Limited survival follow-up - Different mortality endpoints (heterogeneity) - Retrospective analysis vs. prospective application

3.4 Discussion

In this Chapter, a systematic review has identified and evaluated the methodology of risk prediction models that were developed or validating in an IPD meta-analysis. The findings inform current standards, gaps in methodology, and future research needs. The review found 15 relevant articles, all of which developed a model and 11 also undertook some form of validation. The publication dates for these articles ranged from 1994-2008. The key findings were as follows. Seven of the articles used a literature review approach to identify studies, of which six did not obtain IPD for all studies desired; this raises concern of availability bias [100]. The percentage of IPD obtained ranged from 50% to 100%. Missing data was mentioned by 11 of the 15 articles, where three articles deleted some studies due to an absence of a predictor or outcome of interest. Of key importance was the finding that 10 of the 15 articles have developed their model using a ‘one-step meta-analysis ignoring clustering’, this is an unstratified approach, and this may be inappropriate [102]. All 15 articles developed their model using all of the available IPD, but only 11 validated their model, with 9 only using an internal validation approach. Thus only two articles used some form of external validation, via either internal-external cross-validation approach and/or a fully external validation approach. Also, although ten articles provided discrimination statistics, only three provided calibration statistics, and thus calibration performance is rarely evaluated.

Risk prediction models have the potential ability to inform strategies for disease prevention, care, early diagnosis and patient counselling [11]. Their use should be evidence based as with all clinical practice. There must be consistent evidence that a model is reliable and applicable to the intended populations of individuals [16]. This preferably requires the model to be

externally validated in multiple datasets successfully. This can take years to achieve, but with growing access to IPD from large databases, there is an increasing opportunity to develop and validate risk prediction models simultaneously.

This review has allowed an evaluation of current practice and applied statistical methods, identified common methodological challenges, and flagged limitations in current reporting and methodology. These findings will thereby inform those wishing to develop and/or validate a model using IPD from multiple studies, and recommendations are provided below to enhance this. Firstly, the review limitations are discussed.

3.4.1 Limitations of the review

This review has some limitations. Firstly, it only covers articles published up to 2009 which was a restriction due to the database used. However, the key findings are unlikely to be altered if the review were updated; as it was felt qualitative saturation was achieved with this sample in regard the methods used and areas for improvement. However, a second limitation is that, by evaluating published articles, the findings are clearly dependent on the reporting standards within the articles, thus any research deficiencies or methodological gaps may reflect poor reporting (as absence of reporting is not absence of performing). For example, only three articles have referenced a protocol for the IPD project; however this absence of reporting does not necessarily mean the other 12 articles did not have a protocol.

Thirdly, the focus in this review was on articles that utilised IPD from multiple studies, and did not consider a single database containing clusters (e.g. practices). Such articles have similar issues, but the focus was on a typical IPD meta-analysis scenario where IPD studies are obtained and synthesised. Such articles are likely to have similar methods and standards, but this review cannot quantify that.

3.4.2 Methodological challenges to consider

The review identified some key areas for improvement and methodological challenges for researchers to consider when planning to develop/validate a prediction model using IPD from multiple studies. These are summarised in Table 3.7. A major finding is that often clustering of patients within studies is ignored during model development, and this may affect the model estimates and model performance [102]. Dichotomisation/categorisation of predictors is also often performed, and may have resulted in poorly-fitted models and loss of information. There was often a lack of standardisation in data collection between studies, and heterogeneity in predictor effects was often not measured or accounted for (when data is pooled) in the analysis. Missing data was also an issue, with multiple variables often not recorded within each dataset, leading to multiple imputation or omission of prediction variables.

The review has highlighted that those articles that perform a literature review do not always obtain all the IPD that they desire, as some studies are unwilling to provide their data for various reasons. There needs to be a better assessment of the quality of IPD identified, and indeed assessment of study quality as a whole is currently rather neglected. The issue is of bias and whether there is something different about those studies for which IPD can't be obtained.

Table 3.7: Methodological challenges, extending those identified for prognostic factors by Abo-Zaid et al.

Methodological challenges
<ul style="list-style-type: none"> - Identifying relevant studies - Unavailability of IPD in some studies[100] - Issues within studies - How to assess quality of studies identified - Inability of IPD to overcome deficiencies of original studies, such as missing participant data or of being low methodological quality, etc. - Heterogeneity - Different definitions of disease or outcome, e.g. Noordzij et al. [94] note different definitions of hypocalcaemia across studies - Different (or out-dated) treatment strategies, especially when a mixture of older and newer studies are combined; e.g. Yap et al. [99] state that a large proportion of their patients in their included trials did not receive a common post-myocardial infarction therapy - Statistical issues for meta-analysis and model development - Missing data, including: missing values and outcome data for some participants within a study, and unavailable factors in some studies - Difficulty in using a continuous scale for continuous factors in meta-analysis when some studies give IPD values on a continuous scale and others do not (see Rovers et al. [95]) - Whether to account for clustering or not - Examining and incorporating heterogeneity between-study for predictors / alpha term - Assessment of potential biases - How to assess the impact of excluded studies who did not provide IPD [100] - Model validation - Sample size required for internal-external approach

3.4.3 Areas for improvement - Recommendations

This review has highlighted areas for improvement and subsequently recommendations have been provided in Table 3.8 for how to improve the conduct and reporting of future research in this field.

Table 3.8: Areas for improvement and recommendations

Areas for improvement and recommendations
<i>Rationale and initiation</i> <ul style="list-style-type: none"> - Produce a protocol for the project, detailing rationale, conduct and statistical analysis and reference this
<i>Obtaining IPD</i> <ul style="list-style-type: none"> - Report how the primary study authors were approached for their IPD - Report search strategy used, i.e. literature review / collaborative group - Report strategies used for searching the literature for relevant studies (if applicable), including search keywords and databases - Provide a flowchart showing the search strategy, classification of identified articles, and retrieval of IPD from relevant studies - Report sample size requirements and how the amount of IPD was decided upon
<i>Details of IPD</i> <ul style="list-style-type: none"> - Report the number of patients and events for each study used in model development and / or validation - Report the missing data for each study (whether excluded entirely or some data is missing within a study) - Detail the reasons why IPD was unavailable (if applicable), if available then report the number of patients and events from those studies - Compare and report the quality of IPD obtained for each study
<i>Statistical methods for model development</i> <ul style="list-style-type: none"> - Account for clustering of patients within studies, (do not merge all the IPD as if it was one study), using a one-step accounting for clustering approach - Assess and report any between study heterogeneity in the predictors / alpha term (if applicable) - If large heterogeneity does exist, then try to reduce by including more variables or remove heterogeneous variable - Report selection criteria and procedure used to decide inclusion of a predictor in the model - Keep predictors continuous whilst developing a model, unless it is important to categorise with good clinical or statistical reason - Report the final developed model in original format with alpha (if applicable) and beta-estimates, i.e. not just HR, OR, RR - Detail the missing data and variables in each study and how this was dealt with
<i>Assessment of publication and availability biases</i> <ul style="list-style-type: none"> - If applicable, consider the potential impact of publication bias and / or availability bias if studies are not providing IPD, by comparing those studies to the studies providing IPD [100]
<i>Model validation</i> <ul style="list-style-type: none"> - Validate the model that has been developed using both internal and external IPD where possible - If all IPD is desired for model development, perform internal-external cross-validation as recommended by Royston et al. - Report the number of studies used in validating the developed model - Report the number of patients and events used from each study towards validation - Explain the choice of intercept (baseline hazard) to be used when implementing the model in practice - Report validation statistics both in each study and overall

Two important recommendations are discussed here now. The first recommendation is to account for clustering. This is very important because most of the articles did not account for the clustering of patients within different studies, and this can make a difference when interpreting the results for individual studies, because (i) ignoring clustering can bias model estimates [102], and (ii) a model may perform adequately on average but may not do as well for specific populations, for which setting-specific intercepts may be required. This is also highlighted in Chapter 5. The second recommendation is to include some form of external validation where possible; most of the articles did not do this, even though multiple studies are available, as all data was used for model development and then usually internal validation. As multiple studies are available and it may be hard to obtain further data for external validation post meta-analysis, it is imperative to at least consider the use of internal-external cross-validation, as it offers a more robust method of assessing validation outside the model development data.

3.4.3.1 Recommendation One: Allow for different baseline risks in each of the IPD studies

In this review, 10 of the 15 articles did not account for clustering of patients within different IPD studies and therefore their developed prediction model did not allow for any study differences in baseline risk. Although such models can still perform adequately on average (across all studies combined), when applied in practice to specific populations the model performance may not be as good if the population's baseline risk is very different from the average estimated baseline risk. This means that the developed model may require re-calibration in specific populations. Statistically it is

known that omission of an important predictor (i.e. study) can lead to biased effect estimates and reduced power [82]. To address this, Debray et al. [103] recommended the prediction model should be developed with a separate intercept (baseline risk) per study, and then the model's performance can be examined using internal-external validation alongside a strategy for choosing the intercept upon application to the excluded study. Such strategies may include using external knowledge of the intercept in the excluded study population; using the intercept as estimated from the IPD from the excluded study; or taking the intercept estimated from a study used in the model development that contains a similar population to the excluded study. The latter strategy is recommended within Steyerberg et al. [98], where they propose others apply their model using the intercept for one particular trial in their analysis, as this trial reflects the population it is intended for.

Where the intercept can be well-matched to the excluded study, Debray et al. [103] show that their framework allows an IPD meta-analysis to produce a single, integrated prediction model that can be implemented in practice and has improved model performance and generalisability. This echoes other recommendations to account for clustering in an IPD meta-analysis [102]. For survival data, this means that the baseline hazard should be modelled during model development and so researchers should move away from using Cox regression (a common approach used by the articles in the review, but one which does not estimate the baseline hazard) and rather use other approaches such as flexible parametric methods, like the Royston-Parmar model that estimates the baseline hazard using restricted cubic splines [11, 18].

3.4.3.2 Recommendation Two: Implement a framework that uses internal-external cross-validation

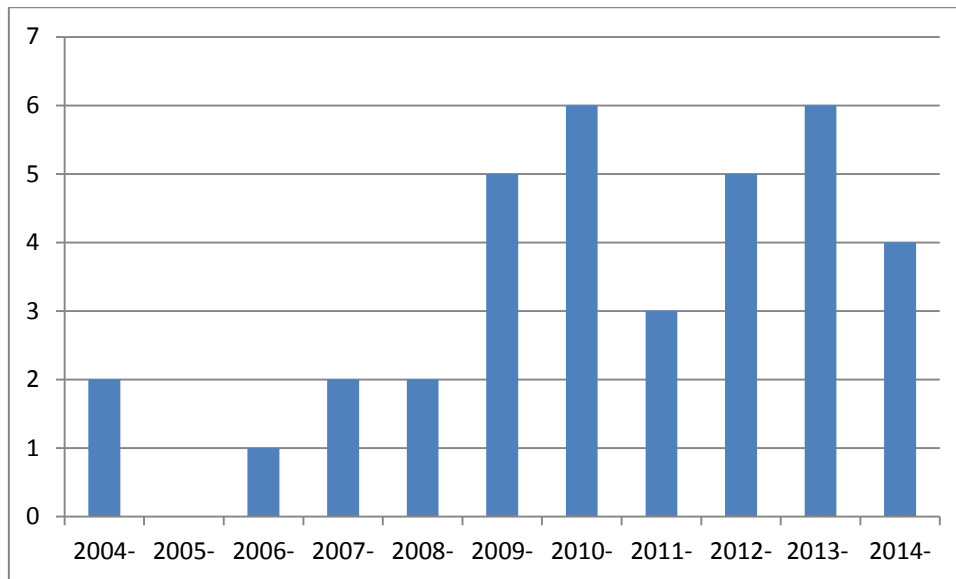
An important finding from this review is that, although multiple studies are available, most researchers develop their model by using IPD from all available studies, thus performing an internal validation (on the same data used to develop the model) rather than an external validation (different data). Only two of the 15 articles performed a form of external validation, and therefore most models require further validation to assess true performance. A possible explanation why researchers choose not to use IPD for external validation is to maximise the data available for model development, which is understandable given the large number of predictors and possible non-linear relationships. Also, if researchers were to consider holding back some IPD for external validation, there is no easy way to decide how much IPD (number of studies) should be held for validation.

However, external validation should be performed if possible as this is ultimately the gold standard for model validation, with an appropriate dataset. For those researchers that wish to use all their data for model development, then this can be achieved alongside some form of external validation, using the internal-external cross-validation approach [101] used by Steyerberg et al. [98] and Yap et al. [99] in the articles reviewed. This approach involves removing one study from the development phase of the model, then fitting the model on the remaining IPD, and then testing performance in the excluded study. The framework is then repeated by rotating the omitted study and assessing validation in all the possible scenarios. Thus model estimates are based on the majority of IPD in each cycle, and the models fit and predictive ability can be compared

across all possible combinations of omitted study. If performance looks adequate, a final step can be to utilise IPD from all studies to provide the final specification of the model. In scenarios where the model fit appears to be poor in some studies, this can indicate a lack of generalisability of the model, highlighting which populations the model is not appropriate for.

Internal-external validation may not always be possible, for example if studies have small sample sizes. But it appears to be an ideal approach to utilise as much IPD for both model development and validation. The article by Royston et al. [101] proposing the internal-external cross-validation approach has been cited 35 times [16, 82, 98, 99, 104-134] between 2004-2015 as of 01/01/2015, by articles that investigate prognostic factors, developing risk prediction models, updating prognostic factor studies and externally validating them, and articles looking at multiple imputation methods and three theses. Nine articles [98, 99, 104, 107, 109, 113-115, 124] have used the internal-external cross-validation method to develop their risk prediction / prognostic model. Figure 3.4 shows the number of citations per year for the Royston et al. [101] paper.

Figure 3.4: Number of articles that have cited the Royston et al. paper, by each year



3.4.4 Conclusions

It is important to consider the statistical and methodological issues when planning to develop and/or validate a risk prediction model from multiple datasets, in order to avoid models that perform poorly and are not generalisable. This review highlights that the IPD meta-analysis approach is very appealing, as it allows the use of internal-external cross validation to develop a model and at the same time evaluate its performance across multiple populations. However, an IPD meta-analysis does not solve all the problems, as numerous challenges remain, particularly missing data and heterogeneity in study quality and methods of measurement. Perhaps an ideal way forward is a prospective IPD meta-analysis, where researchers agree at their study onset to use set quality standards and record particular variables in a common way so that, upon their own study completion, they can supply their IPD to those developing/validating a risk prediction model. Heterogeneity can then be limited by researchers agreeing before data

collection to standardize predictor definitions, measurement methods, and outcome recoding.

3.4.5 Findings in context for the thesis

This review has been published in *BMC Medical Research Methodology* [16]. The findings raise many methodological challenges, and the next two chapters are motivated by one of the key findings: the issue of clustering of studies, and whether there is a difference in outcome prediction when accounting for between study heterogeneity. In particular, articles such as Noordzij [94] do not account for clustering of studies, and it is of interest if and how this affects absolute risk predictions.

3.4.6 What this chapter adds?

A short summary of this chapter in terms of what the problem is, what this chapter adds and what is needed next, is provided in Table 3.9.

Table 3.9: What chapter three adds?

What is known / what is the problem?
<ul style="list-style-type: none"> • Important to do a systematic review of articles using IPD to develop and validate risk prediction models • Assess whether the issues found in Chapter 2 are prevalent in the wider research community
What this study adds?
<ul style="list-style-type: none"> • A systematic review such as this has not been performed before for studies that use IPD • This study highlights an important issue, where multiple studies are used to develop a model but their heterogeneity is not accounted for and IPD is instead regarded as coming from one study • A key contribution of this work is that more researchers should consider clustering • This study has found that models are usually internally validated and rarely externally validated, with a few articles recently using an approach called internal-external cross-validation • Internal-external cross-validation is a novel approach that more researchers should be thinking about and helps to get a better assessment of the generalisability of the model
What is needed next?
<ul style="list-style-type: none"> • In the next two chapters, discrimination and calibration will be assessed for an already published set of results for a risk prediction model found through this review • The results will be compared to the standard approach of not accounting for heterogeneity • As well as assessing internal-external cross-validation

CHAPTER 4: UNSTRATIFIED VERSUS META-ANALYTIC APPROACHES FOR EVALUATING PREDICTIVE TEST ACCURACY USING IPD: PART 1 - DISCRIMINATION

4.1 Introduction and objectives

The previous chapter reviewed risk prediction model projects that used IPD from multiple studies, and found that many ignore clustering and treat IPD as if it is all from one study. Here, one of those articles is taken and the implications of ignoring clustering are explored. The aim of this chapter is to illustrate the benefits of having IPD for meta-analysis of risk prediction studies, and to compare how an unstratified analysis that treats all IPD as coming from a single study compares to a more sophisticated analysis that accounts for clustering using more sophisticated meta-analysis methods. The focus is on a single predictive test (in other words, a risk prediction model that contains a single predictor) and its discrimination ability when using unstratified and meta-analysis approaches. Discrimination is the ability of a test or prediction model to differentiate high risk patients from low risk patients [135], as introduced in Chapter 1. Calibration is considered separately in Chapter 5.

The chapter uses a dataset from a paper identified in Chapter 3 through the systematic review. This paper is Noordzij et al. [94] who use IPD from nine observational studies to predict hypocalcaemia after thyroidectomy using parathyroid hormone (PTH). This chapter begins with a summary of this article and then introduces, applies and compares meta-analytic approaches that do or do not account for clustering.

The discrimination and calibration work performed in Chapters 4 and 5 has been published in *Statistics in Medicine* (Riley et al. [42]), where I am the second author. PTH has been used as a case study within the paper and my analysis for discrimination and calibration is used as illustrated examples; I also contributed heavily to writing these sections within the paper and drafting and finalising the paper.

4.2 Summary of Noordzij article

Noordzij et al. [94] conducted a systematic search for papers which describe the use of parathyroid hormone (PTH) assay (a continuous variable) to predict postoperative hypocalcaemia after thyroidectomy. To be included, a study had to record PTH assay checked within hours of completing thyroidectomy and link this to a subsequent diagnosis of hypocalcaemia defined by low calcium levels. A thyroidectomy is an operation that involves the surgical removal of all or part of the thyroid gland [136]. The issue is that monitoring for hypocalcaemia after thyroidectomy using only serum calcium levels can delay the discharge of patients who will remain normocalcaemic patients, and also delay the treatment of hypocalcaemia patients [94]. Thus earlier identification of those at high risk of hypocalcaemia would help decide which patients not to send home, and conversely which patients are low risk and could be discharged.

Hypocalcaemia is a condition where low serum calcium levels are present in the blood [137, 138]. Calcium is vital for the development of healthy bones and teeth. It is also needed for muscle contraction, heartbeat regulation and formation of blood clots. A calcium deficiency can lead to brittle-bone disease. The parathyroid glands are responsible for regulating the calcium levels in the body. When calcium levels drop the parathyroid glands secrete PTH. PTH causes calcium to be released from bones, so more can be reabsorbed by the kidneys

and also from food in the intestines. Low levels of calcium can cause muscular spasms and numbness in the hands, feet, mouth and even throat, as well as depression and seizures. In the case of hypocalcaemia, vitamin D and calcium supplements are needed. Hypocalcaemia may be caused by chronic kidney failure, pancreatitis or problems with parathyroid glands which means they do not produce enough PTH. This often occurs after damage to the parathyroid glands during surgery on the thyroid glands. PTH is measured through a sample of a patients' blood.

Noordzij et al. 2007 [94] performed a medical literature search in PubMed using the keywords: 'PTH' (parathyroid hormone) and 'thyroidectomy' for all English articles that were published between 1996 and January 2006. They included studies if PTH was obtained within 24 hours of completing total or completion thyroidectomy, and only included observational studies. A total thyroidectomy is when the entire thyroid gland is removed [139], and completion thyroidectomy involves removing the remaining thyroid tissue after the patient has previously had partial removal [140]. Noordzij et al. [94] excluded studies if all the patients were treated with postoperative calcium, or if early PTH values were used to change how the patient was managed. Fifteen papers met the inclusion criteria and the authors of those 15 papers were e-mailed to request 'their' IPD, of which only 9 provided it.

IPD containing preoperative PTH and calcium levels, and whether they developed hypocalcaemia, were obtained for 457 patients from the 9 studies and Noordzij et al. [94] pooled these to yield the unstratified results across the studies. PTH was checked up to three time periods after removal of the thyroid gland (0 to 20 minutes, 1 to 2 hours, and 6 hours); PTH levels were found to be substantially lower in patients who became hypocalcaemic compared with those who remained normocalcaemic. The accuracy of PTH in determining

hypocalcaemia increased with time, and was best when checked from 1 to 6 hours postoperatively. IPD, including original PTH values, enabled % change in PTH from baseline to be calculated. This was used by the authors, rather than the absolute change in PTH, due to different methods of measuring PTH across studies. A single PTH threshold (65% decrease in PTH compared with preoperative level), checked 6 hours after completing thyroidectomy had a sensitivity of 96.4% and specificity of 91.4% in detecting postoperative hypocalcaemia, and was recommended by the authors as the optimal threshold to detect hypocalcaemia.

Noordzij et al. [94] conclude from their analysis of the data that routine use of this assay should be considered because it may allow earlier discharge of the normocalcaemic patient and earlier identification of patients requiring treatment of post-thyroidectomy hypocalcaemia [94]. However, a potential limitation of their work is their analysis, which has merged all the IPD together and ignored both clustering of patients within studies (unstratified) and heterogeneity of test accuracy across studies.

Noordzij kindly provided the data for the purposes of this thesis. In this chapter, their original analysis has been replicated using the original thresholds used, and then extended to properly model clustering and heterogeneity. The results of the different approaches are then compared. Noordzij [94] dichotomised PTH into high and low using a threshold, and examined multiple choices for the cut off. To be consistent this has also been done in this thesis. Possible extensions for modelling PTH on its continuous scale are discussed in the Discussion, Chapter 7.

4.3 Summary of the IPD available

The 6 IPD studies of Noordzij are summarised in Table 14. Although there was data for 457 patients from the full set of nine studies, three of these studies were eventually excluded for not providing pre-operative values, which is why in Table 4.1 the overall number of patients in the reported articles is reduced down to 388.

Table 4.1: Summary of the 6 IPD studies

Study	No of patients in reported article	No of events in reported article	Outcome Prevalence in reported article	Prospective?	Definition of hypocalcaemia (equivalent cCa (mg/dL))	IPD provided: no. of patients (events)	Outcome Prevalence in IPD	Times at which PTH was measured
Lam 2003 [141]	40	12	0.30	Yes	<7.6	39 (12)	0.31	Pre-op, 1 and 6 hour post-op
Lo 2002 [142]	155	32	0.21	Yes	<7.2	100 (11)	0.11	Pre-op and intra-op
Lombardi 2004 [143]	53	16	0.30	Yes	<8.0	52 (16)	0.31	Pre-op, intra-op, 2, 4, 6, 24 and 48 hour post-op
McLeod 2006 [144]	60	15	0.25	Yes	<8.0	60 (14)	0.23	Pre-op, intra-op and 1 hour post-op
Warren 2002 [145]	53	5	0.09	Yes	<8.0	22 (4)	0.18	Pre-op, Intra-op and 1 hour post-op
Warren 2004 [146]	27	3	0.11	Yes	<8.0	27 (3)	0.11	Pre-op, intra-op and 1 hour post-op
Overall	388	83	0.21			300(60)	0.20	

The number of patients reported differs throughout each study as Table 4.1 shows, from 27 to 155, with a mean of 64.7. All six studies are prospective studies and the definition of hypocalcemia ranges from <7.2 mg/dL to <8.0 mg/dL, with 4 studies using the value of <8.0 mg/dL. So the reference standard is fairly consistent across studies. The IPD used by Noordzij et al. [94] differs from the number of patients reported in the original studies as there are data missing for 88 patients. No reason is given by Noordzij [94] as

to why a reduced number of patients was given in the IPD provided by 4 of the 6 of the studies.

Note also that the prevalence of the outcome varies across studies, from 0.09 to 0.30 (as reported in the original articles) and 0.11 to 0.31 (in the IPD provided), potentially due to the different clinical settings and patient groups included in each study. Noordzij et al. [94] have used the following thresholds: 40%, 50%, 60%, 65%, 70%, 72%, 80% and 90%. For consistency, the same thresholds will be used in this analysis. Discrete thresholds are being used because clinical decisions will use thresholds, and I wanted to compare directly to the original paper by Noordzij, who focused on thresholds.

4.4 Methods

The IPD analysis in this chapter aims to compare statistical approaches for examining the discriminatory ability of PTH, for a variety of thresholds for % change from baseline, when measured 0-20 minutes, 1-2 hours and 6 hours from end of surgery. The methods are now described.

Initially t-tests are used as a preliminary analysis to compare mean values, as this has also been performed by Noordzij et al. [94] to summarise the separation between those with and without the outcome. The following two approaches are used:

- i) t-test at each time point in the unstratified dataset
- ii) t-test repeated for each study separately

In both of these approaches, the mean PTH values are calculated in the two groups hypocalcemia and normocalcemia (patients who do not have hypocalcemia) and p-

values are calculated which indicate whether there is a difference in the means of these two groups, a p-value <0.05 is regarded as being statistically significant. The difference in the mean PTH values is calculated for pre-operative values, 0-20 minutes, 1-2 hours and 6 hours.

However, discrimination is better considered in terms of sensitivity, specificity and the C statistic (area under the curve). Different meta-analysis approaches are now described for obtaining a summary sensitivity, specificity and C statistic across studies. Firstly, the unstratified analysis approach of Noordzij et al. [94], followed by some alternative meta-analysis approaches that properly account for clustering and heterogeneity.

4.4.1 Noordzij analysis: unstratified approach

In the unstratified approach, IPD from all studies are combined and treated as one dataset, with clustering and heterogeneity ignored. Sensitivity and specificity for a range of thresholds, and the C statistic is calculated for percentage PTH decrease (percentage decrease from pre-op level to either: 0-20 minutes, 1-2 hours or 6 hours).

% PTH decrease

% PTH decrease is calculated by:

$$\% \text{ PTH decrease} = ((\text{preoperative PTH} - \text{post operative PTH}) / \text{preoperative PTH}) \times 100$$

(Equation 4.1)

Then using a threshold value, the IPD is summarised by:

Table 4.2: Two by two table for each threshold

	Hypocalcaemic	Normocalcaemic
Positive (> threshold %)	a	b
Negative (≤ threshold %)	c	d

Where a is the true positive b is the false positive, c is the false negative and d is true negative

Sensitivity

Sensitivity in the unstratified analysis is calculated using:

$$\text{Sensitivity} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}}$$

= probability of a positive PTH test given that a patient becomes hypocalcaemic

Also, from a 2 by 2 table (Table 4.2):

$$\text{Sensitivity} = \frac{a}{a + c}$$

(Equation 4.2)

Specificity

Specificity in the unstratified analysis is calculated using:

$$\text{Specificity} = \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}}$$

= probability of a negative PTH test given that the patient does not
become hypocalcaemic

Also, from a 2 by 2 table (Table 4.2):

$$\text{Specificity} = \frac{d}{b + d}$$

(Equation 4.3)

Youden's J statistic

This is a statistic that is used to determine the performance of a diagnostic test. This was suggested by Youden [147] to summarise the performance of a test, a value of 0 indicates the test is useless and a value of 1 indicates the test is perfect. This statistic gives equal weighting to false positives and false negatives. It is applicable to a predictive test such as this which is being investigated and will be used to determine the best threshold. This will be multiplied by 100 as sensitivity and specificity have been multiplied by 100 as well.

$$J = \text{Sensitivity} + \text{Specificity} - 1$$

(Equation 4.4)

C statistic

To calculate the C statistic, sensitivity versus 1-specificity was plotted on an ROC curve and then integrated using the trapezoid function to get the area under the curve [148]. The trapezoid function integrates trapeziums under the curve, to work out their area and then add them up to give the AUC, with points at (0, 0) and (1, 1) added manually to give the full AUC. To plot an ROC curve for continuous data for each time-point, the ROCTAB command was used in STATA 12.1 [149]. ROCTAB gave a C statistic with its standard error and a 95% confidence interval.

4.4.2 Meta-analysis approaches

Rather than regarding the data as coming from one study only, a meta-analysis that accounts for clustering and potential between-study heterogeneity is also possible. The meta-analysis models are now explained.

4.4.2.1 Univariate meta-analysis for sensitivity and specificity

A univariate meta-analysis of sensitivity is:

$$r_{+ve_i} \sim \text{Bin}(\rho_{\text{Sens}_i}, N_{\text{diseased}_i})$$

$$\text{Logit}(\rho_{\text{Sens}_i}) \sim N(\theta_{\text{Sens}}, \tau_{\text{Sens}}^2)$$

(Equation 4.5)

Where r_{+ve_i} is the number of true positives, ρ_{Sens_i} is the true sensitivity in study i , N_{diseased_i} represents the total number with hypocalcaemia, θ_{Sens} is the average logit-sensitivity across studies, τ_{Sens}^2 is the between-study heterogeneity in the logit-sensitivity.

For specificity:

$$r_{-ve_i} \sim \text{Bin}(\rho_{\text{Spec}_i}, N_{\text{non-diseased}_i})$$

$$\text{Logit}(\rho_{\text{Spec}_i}) \sim N(\theta_{\text{Spec}}, \tau_{\text{Spec}}^2)$$

(Equation 4.6)

Where r_{-ve_i} is the number of true negatives, ρ_{Spec_i} is the true specificity in study i , $N_{\text{non-diseased}_i}$ represents the total without hypocalcaemia, θ_{Spec} is the average logit-specificity across studies, and τ_{Spec}^2 is the between-study heterogeneity in the logit specificity.

The meta-analysis models utilise the exact binomial distribution within each study, and estimate the average sensitivity and specificity across studies[150, 151], and the variability of sensitivity and specificity across studies.

4.4.2.2 Bivariate meta-analysis for sensitivity and specificity

In this bivariate meta-analysis approach recommended by the Cochrane Screening and Diagnostic Test Methods Group [22], sensitivity and specificity are jointly analysed to account for any between-study correlation as follows:

$$\begin{aligned}
 r_{+ve_i} &\sim \text{Bin}(\rho_{\text{Sens}_i}, N_{\text{diseased}_i}) \\
 r_{-ve_i} &\sim \text{Bin}(\rho_{\text{Spec}_i}, N_{\text{non-diseased}_i}) \\
 \begin{matrix} \text{Logit}(\rho_{\text{Sens}_i}) \\ \text{Logit}(\rho_{\text{Spec}_i}) \end{matrix} &\sim N \left(\begin{matrix} \theta_{\text{Sens}} \\ \theta_{\text{Spec}} \end{matrix}, \begin{bmatrix} \tau_{\text{Sens}}^2 & \tau_{\text{Sens,Spec}} \\ \tau_{\text{Sens,Spec}} & \tau_{\text{Spec}}^2 \end{bmatrix} \right)
 \end{aligned}$$

(Equation 4.7)

In this model, the parameters are as defined previously, with the addition of $\tau_{\text{Sens,Spec}}$ which is the between-study covariance in logit sensitivity and logit specificity.

Covariance might arise because both sensitivity and specificity are being estimated together, so there may be some correlation between them. Usually this is caused by changes in the threshold across studies, but here a common threshold is used and so the correlation should be close to 0 and thus the univariate approach may be sensible.

The XTMELOGIT command is used in STATA 12.1 [149] to perform both univariate and bivariate meta-analyses, and STATA 12.1 [149] is used for all other analyses performed in this chapter. Maximum likelihood estimation method is used and numerical integration. Different numbers of quadrature points can be specified, as the number of points increases so does the estimation accuracy but this also increases the computational time. Five quadrature points were chosen for this analysis as it gave

estimates close to those when using >10 points but with a much faster computational time.

4.4.2.3 C statistic

There are a number of ways that a summary C statistic could be derived from a meta-analysis. The aforementioned bivariate or univariate meta-analyses can be done for each threshold separately, to give a summary of sensitivity and specificity at each, enabling a summary ROC curve to be plotted and the area under it gives a summary C statistic.

Alternatively, and perhaps preferably, a C statistic can be obtained for each study separately and then pooled to give a summary C statistic. This is the approach taken here, and the C statistic estimates are combined in a random-effects meta-analysis (see 1.4.2.4) to obtain an average C statistic [27], with the C statistic analysed on its original scale as recommended by Klaveren et al. [27]. The random-effects model can be written as:

$$c_i \sim N(\theta_i, s_i^2)$$

$$\theta_i \sim N(\mu_c, \tau_c^2)$$

(Equation 4.8)

Where c_i is the C estimate in study i and s_i^2 is its variance, which is assumed to be known; θ_i if the true C statistic in study i , which is assumed to be from a normal distribution with a mean of μ_c and between-study variance τ_c^2 .

A 95% prediction interval for this C statistic, which gives a range for the predicted value of the C statistic in a new study, is approximately:

$$\hat{\mu} - t_{k-2}\sqrt{\hat{\tau}^2 + SE(\hat{\mu})^2}, \hat{\mu} + t_{k-2}\sqrt{\hat{\tau}^2 + SE(\hat{\mu})^2}$$

(Equation 4.9)

Where $\hat{\mu}$ is the estimate of the average C statistic across studies from the random-effects meta-analysis, $SE(\hat{\mu})$ is the standard error of $\hat{\mu}$, $\hat{\tau}$ is the estimate of between-study standard deviation, t_{k-2} is the $100(1 - \alpha/2)$ percentile of the t-distribution with k-2 degrees of freedom, where k is the number of studies in the meta-analysis, and α is usually chosen as 0.05, to give a 5% significance level and thus 95% prediction interval [32].

4.5 Results

In this section, presented and compared are the results of the unstratified, univariate meta-analysis and bivariate meta-analysis approaches described in 4.4. There are three comparisons (i) t-test analysis, (ii) sensitivity and specificity, and (iii) C statistic, with each done at every time-point that PTH was measured.

Firstly, Table 4.3 summarises the data available, and the t-test results for any difference of mean PTH values in hypocalcaemic and normocalcaemic patients.

Table 4.3: Summary of the PTH values before and after thyroidectomy pooled across the 6 studies and in each study separately

Patients	Preoperative	P-value $H_0:d=0$ $H_a:d \neq 0$	0-20 minutes	P-value $H_0:d=0$ $H_a:d \neq 0$	1-2 hrs	P-value $H_0:d=0$ $H_a:d \neq 0$	6 hrs	P-value
All	59.57 ± 31.51 (n=292)		29.91 ± 24.35 (n=256)		28.52± 24.41 (n=169)		24.80± 20.23 (n=84)	
Normocalcaemic	60.48 ± 32.37 (n=232)	P=0.17	34.24 ± 24.16 (n=209)	P<0.0001	36.08± 23.66 (n=124)	P<0.0001	34.41± 17.59 (n=56)	P<0.0001
Hypocalcaemic	56.03 ± 28.35 (n=60)		10.66 ± 13.56 (n=47)		7.70± 10.32 (n=45)		4.52 ± 3.31 (n=28)	
McLeod[144] Normocalcaemic	75.81 ± 38.93 (n=46)	P=0.35	37.57 ± 21.54 (n=43)	P=0.0002	36.76 ± 17.45 (n=25)	P<0.0001		
McLeod[144] Hypocalcaemic	71.41 ± 24.59		13.97 ± 11.23		8.28 ± 5.71			

Patients	Preoperative	P-value $H_0:d=0$ $H_a:d\neq 0$	0-20 minutes	P-value $H_0:d=0$ $H_a:d\neq 0$	1-2 hrs	P-value $H_0:d=0$ $H_a:d\neq 0$	6 hrs	P-value
Warren 2002[145] Normocalcaemic	(n=14) 50.41 ± 17.43	P=0.63	(n=13) 38.47 ± 28.04	P=0.12	(n=12) 41.93 ± 36.26	P=0.39		
Warren 2002[145] Hypocalcaemic	(n=12) 54.53 ± 28.83		(n=18) 20.10 ± 26.10		(n=14) 34.00 ± 38.18			
Warren 2004[146] Normocalcaemic	(n=4) 65.41 ± 35.12	P=0.93	(n=4) 37.57 ± 21.54	P=0.26	(n=2) 36.76 ± 17.45	P=0.05		
Warren 2004[146] Hypocalcaemic	(n=23) 99.70 ± 55.15		(n=43) 28.80 ± 32.44		(n=25) 18.33 ± 17.10			
Lam[141] Normocalcaemic	(n=3) 45.56 ± 33.16	P=0.71	(n=3)		(n=3) 33.00 ± 21.24	P<0.0001	36.04 ± 20.00	P<0.0001
Lam[141] Hypocalcaemic	(n=27) 51.83 ± 27.76				(n=26) 3.83 ± 1.95		(n=23) 3.92 ± 2.39	
Lo[142] Normocalcaemic	(n=12) 60.17 ± 29.42	P=0.09	28.16 ± 20.89	P<0.0001				
Lo[142] Hypocalcaemic	(n=89) 47.97 ± 20.26		(n=89) 1.86 ± 3.24					
Lombardi[143] Normocalcaemic	(n=11) 52.87 ± 21.98	P=0.07	(n=11) 35.90 ± 18.32	P<0.0001	32.04 ± 16.65	P<0.0001	33.30 ± 15.99	P<0.0001
Lombardi[143] Hypocalcaemic	(n=35) 43.44 ± 19.80		(n=35) 8.26 ± 4.20		(n=35) 4.87 ± 3.73		(n=34) 5.01 ± 3.91	
	(n=16)		(n=16)		(n=16)		(n=16)	

#P-values from t-test, d represents the difference in means

4.5.1 Comparison of mean values

i) t-test at each time point (unstratified approach)

Two-sample mean comparison test results are shown in Table 4.3 with the p-values next to their corresponding groups that have been tested. When all the IPD is put together and analysed unstratified, a p-value of 0.17 is obtained for the difference between the means of normocalcaemic and hypocalcaemic patients at the preoperative level. This indicates there is not a statistically significant difference present here. For the other three time points, they have all got p-values <0.0001 which indicate there is a statistically significant difference present between the means of normocalcaemic and hypocalcaemic patients. PTH values are lower for hypocalcaemia patients. This result is

encouraging as it suggests that PTH is a potential predictive factor (i.e. low levels are associated with being hypocalcaemic) when measured post-surgery.

ii) t-test repeated for each study separately

When looking at each study separately for the preoperative level (Table 4.3), none of the six studies which have the preoperative PTH levels have a statistically significant p-value for difference of means between normocalcaemic and hypocalcaemic patients, although the Lo [142] study ($p=0.09$) and Lombardi [143] ($p=0.07$) have very low p-values that are close to significance.

For the 0-20 minutes time point (Table 4.3): McLeod [144], Lo [142] and Lombardi [143] have a very small p-value that is statistically significant and shows a visible difference in means, whereas the Warren 2002 & 2004 [145, 146] studies do not have statistically significant p-values.

For the 1-2 hour time point (Table 4.3): McLeod [144], Lam [141] and Lombardi [143] all have very small p-values that are statistically significant in showing there is a difference in means of normocalcaemic and hypocalcaemic patients. Warren 2002 [145] study has a $p=0.39$ which is not statistically significant and Warren 2004 [146] has a $p=0.05$ which is borderline significant.

For the 6 hour time point (Table 4.3): Lam [141] and Lombardi [143] studies have all got very small p-values that are statistically significant for difference of means between normocalcaemic and hypocalcaemic patients.

In summary, building on the unstratified analysis of Noordzij et al, these results show that for the majority of studies there is a statistically significant difference between the

post-operative PTH means of normocalcaemic and hypocalcaemic patients. This strongly indicates again that PTH is a prognostic factor, but discriminative ability for individual patients is better summarised by sensitivity, specificity and the C statistic.

4.5.2 Sensitivity and Specificity

4.5.2.1 Comparison of discrimination results at 0-20 minutes

Table 4.4 shows the univariate and bivariate meta-analysis pooled results and confidence intervals are generally very similar to the unstratified analysis for both sensitivity and specificity. However, the confidence intervals are slightly wider in the meta-analysis approaches, as they additionally include the between-study heterogeneity which is often non-zero. Also, at a 90% PTH decrease, the summary sensitivity is much lower in the meta-analysis approaches e.g. the sensitivity is 20.27% in univariate meta-analysis and 34.04% in the unstratified analysis, showing that the unstratified approach can indeed give different summary results. Similarly, specificity is higher in the meta-analysis than the unstratified analysis from the thresholds 60%-72%. The results between unstratified analysis and meta-analysis are fairly similar and only differ substantially at the 90% threshold where there is between-study heterogeneity present (4.33 and 3.41 for univariate and bivariate respectively) in the sensitivity values. In most analyses heterogeneity is close to zero for sensitivity, but above 0.2 for specificity.

Table 4.4: Sensitivity and Specificity of PTH assay in predicting postoperative hypocalcaemia for 0-20 minutes: unstratified, univariate and bivariate results

0-20 minutes	Summary Sensitivity	95% CI		Summary Specificity	95% CI		T^2_{Sens}	T^2_{Spec}	Between-study Correlation
% PTH Decrease		Lower	Upper		Lower	Upper			
Unstratified analysis									
>40	93.62	82.84	97.81	44.06	37.39	50.95	-	-	-
>50	93.62	82.84	97.81	56.44	49.54	63.09	-	-	-
>60	85.11	72.31	92.59	70.79	64.18	76.63	-	-	-
>65	80.85	67.46	89.58	74.26	67.81	79.79	-	-	-
>70	80.85	67.46	89.58	75.74	69.39	81.14	-	-	-
>72	80.85	67.46	89.58	76.73	70.44	82.03	-	-	-
>80	68.09	53.83	79.60	85.64	80.14	89.81	-	-	-
>90	34.04	22.17	48.33	95.54	91.75	97.64	-	-	-
Univariate meta-analysis									
>40	93.62	82.00	97.93	46.75	35.83	58.00	0.00	0.13	-
>50	93.62	82.00	97.93	56.44	49.52	63.11	0.00	0.00	-
>60	85.11	71.91	92.73	75.58	63.36	84.72	0.00	0.20	-
>65	80.85	67.12	89.72	79.05	67.03	87.50	0.00	0.24	-
>70	80.85	67.12	89.72	79.78	68.59	87.70	0.00	0.20	-
>72	80.85	67.12	89.72	81.55	69.71	89.46	0.00	0.27	-
>80	70.21	55.78	81.50	85.15	79.55	89.42	0.00	0.00	-
>90	20.27	2.51	71.53	96.17	89.29	98.69	4.33	0.27	-
Bivariate meta-analysis									
>40	94.00	81.32	98.26	47.01	36.03	58.29	0.09	0.13	-1
>50	93.62	82.00	97.93	56.44	49.52	63.11	0.00	0.00	-0.98
>60	85.19	67.94	93.98	75.86	63.93	84.79	0.34	0.20	-1
>65	81.69	63.22	92.05	79.10	67.71	87.23	0.34	0.21	-1
>70	81.91	62.94	92.36	79.96	69.30	87.58	0.34	0.21	-1
>72	82.21	63.42	92.49	81.62	70.44	89.22	0.37	0.24	-1
>80	70.64	54.78	82.69	85.88	78.87	90.83	0.07	0.03	-1
>90	23.98	3.96	70.70	96.42	91.16	98.60	3.41	0.10	-0.99

Figure 4.1: ROC curves comparing the unstratified and meta-analysis approaches for 0-20 minutes

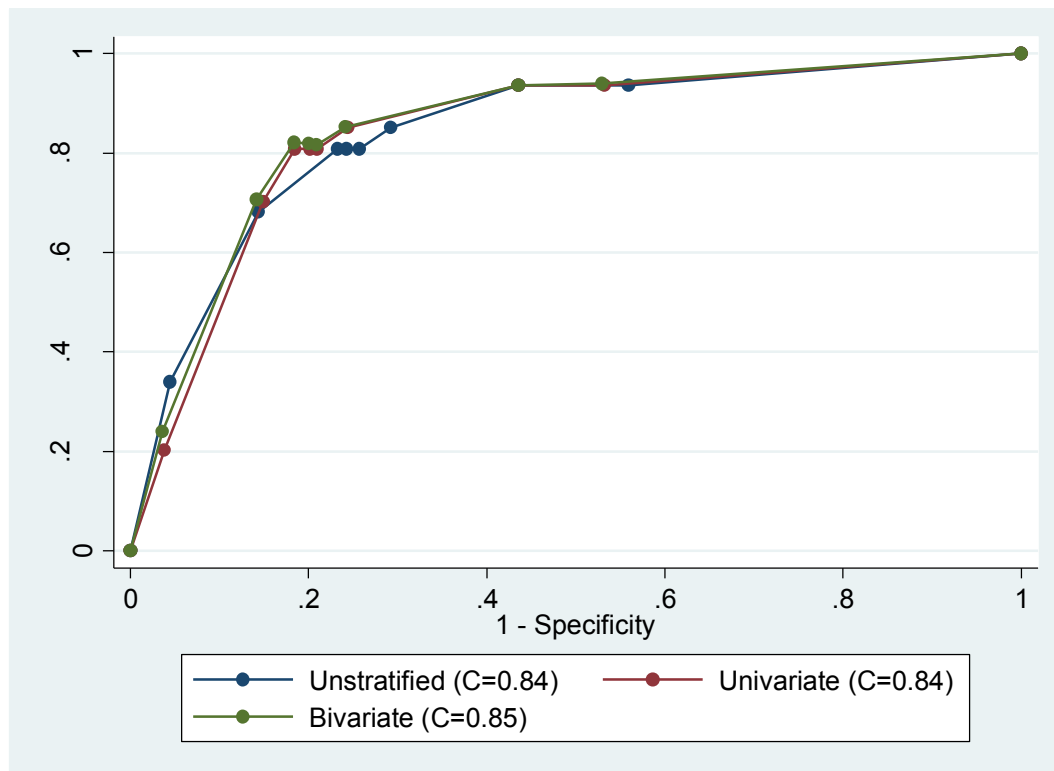


Figure 4.1 illustrates the difference between the three methods is generally small in terms of the actual estimates of sensitivity and specificity. However, the C statistic (AUC) is slightly higher for the bivariate meta-analysis compared to the unstratified and univariate approaches (0.85 compared to 0.84).

In summary of the analysis of PTH at 0-20 minutes post-surgery, differences in unstratified and meta-analysis approaches are often small, but can occur for the summary sensitivity and specificity, especially when there is between-study heterogeneity, which can further impact upon the summary ROC and AUC. Here the unstratified approach is slightly lower in terms of the summary specificity between thresholds 60 to 72, and assumes no between-study heterogeneity when often it exists.

The univariate and bivariate results are generally very similar, but interestingly the correlation in the bivariate models is often poorly estimated at -1 [152]. The bivariate approach is therefore perhaps unnecessarily complex here.

Table 4.5: Table for Youden's statistic comparing the different approaches for 0-20 minutes

0-20 minutes % PTH Decrease	Youden's statistic		
	Unstratified analysis	Univariate meta-analysis	Bivariate meta-analysis
>40	37.68	40.37	41.01
>50	50.06	50.05	50.06
>60	55.90	60.69	61.05
>65	55.11	59.90	60.79
>70	56.59	60.64	61.87
>72	57.58	62.40	63.83
>80	53.73	55.36	56.52
>90	29.58	16.43	20.40

Table 4.5 shows that unstratified and meta-analysis approaches find that Youden's statistic is at its highest for the cut off of 72% PTH decrease, which is encouraging as it agrees with the original authors' conclusion about the best threshold. This assumes sensitivity and specificity are equally important in this clinical setting as Noordzij did, but this may not be true (see Chapter 5, section 5.5).

4.5.2.2 Comparison of sensitivity and specificity at 1-2 hours and 6 hours

Results for sensitivity and specificity using PTH at 1-2 hours and 6 hours are summarised in Tables 4.6 to 4.9, and Figures 4.2-4.3. Generally there is very little difference between the unstratified, univariate and bivariate approaches, as the between-study heterogeneity for all thresholds is very small or zero. The bivariate model again struggles to estimate the correlation, and all methods agree about the optimal choice of threshold.

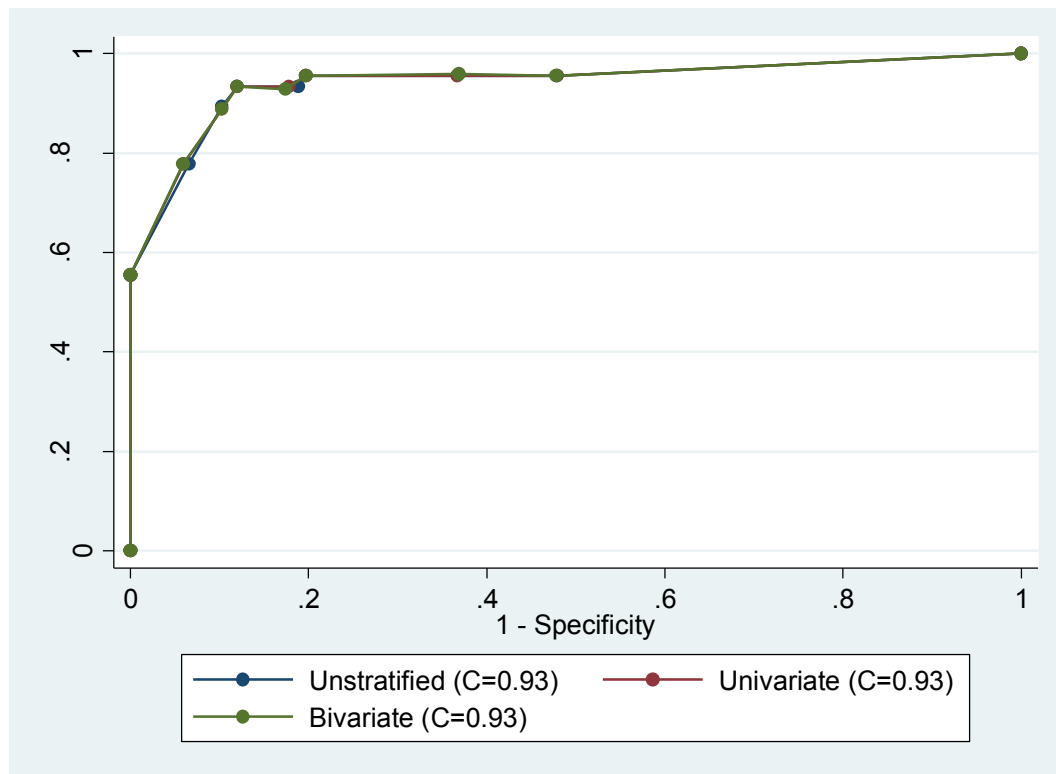
As threshold increases, sensitivity should decrease. However, interestingly in Table 4.6 and Table 4.8, the bivariate results give a larger sensitivity at 50% than 40% (and also

70% is higher than 65% in Table 4.6). This occurs because T^2_{Sens} changes at each threshold and this impacts the pooled result.

Table 4.6: Sensitivity and Specificity of PTH assay in predicting postoperative hypocalcaemia for 1-2 hours: unstratified, univariate and bivariate results

1-2 hours % PTH Decrease	Summary Sensitivity	95% CI		Summary Specificity	95% CI		T ² _{Sens}	T ² _{Spec}	Correlation
		Lower	Upper		Lower	Upper			
Unstratified analysis									
>40	95.56	85.17	98.77	52.14	43.16	60.98	-	-	-
>50	95.56	85.17	98.77	63.25	54.22	71.43	-	-	-
>60	95.56	85.17	98.77	80.34	72.23	86.53	-	-	-
>65	93.33	82.14	97.71	81.20	73.17	87.24	-	-	-
>70	93.33	82.14	97.71	88.03	80.91	92.74	-	-	-
>72	89.36	77.41	95.37	89.74	82.93	94.04	-	-	-
>80	77.78	63.73	87.46	93.46	87.11	96.80	-	-	-
>90	55.56	41.18	69.06	100.0	96.82	100.00	-	-	-
Univariate meta-analysis									
>40	95.56	83.89	98.89	52.14	43.11	61.02	0.00	0.00	-
>50	95.56	83.89	98.89	63.31	53.83	71.85	0.00	0.01	-
>60	95.56	83.89	98.89	80.34	72.15	86.57	0.00	0.00	-
>65	93.33	81.27	97.83	82.19	74.69	88.86	0.00	0.02	-
>70	93.33	81.27	97.83	88.03	80.81	92.78	0.00	0.00	-
>72	88.89	75.95	95.30	89.74	82.80	94.08	0.00	0.00	-
>80	77.78	63.41	87.61	94.02	87.98	97.12	0.00	0.00	-
>90	55.56	40.98	69.24	100.00	Not given	100.00	0.00	0.06	-
Bivariate meta-analysis									
>40	95.60	83.14	98.97	52.23	42.68	61.63	0.05	0.01	-1
>50	95.92	79.58	99.30	63.16	53.00	72.28	0.30	0.03	1
>60	95.56	83.89	98.89	80.34	72.15	86.57	0.00	0.00	0.55
>65	92.90	78.16	97.95	82.59	71.41	90.01	0.15	0.07	1
>70	93.33	81.27	97.83	88.03	80.81	92.78	0.00	0.00	1
>72	88.89	75.95	95.30	89.74	82.80	94.08	0.00	0.00	-1
>80	77.80	63.14	87.77	94.12	87.19	97.42	0.02	0.04	1
>90	55.56	40.98	69.24	100.00	Not given	100.00	0.00	4.66	-0.06

Figure 4.2: ROC curves comparing the unstratified and meta-analysis approaches for 1-2 hours



#The differences are barely visible due to near identical curves, as they are superimposed on each other

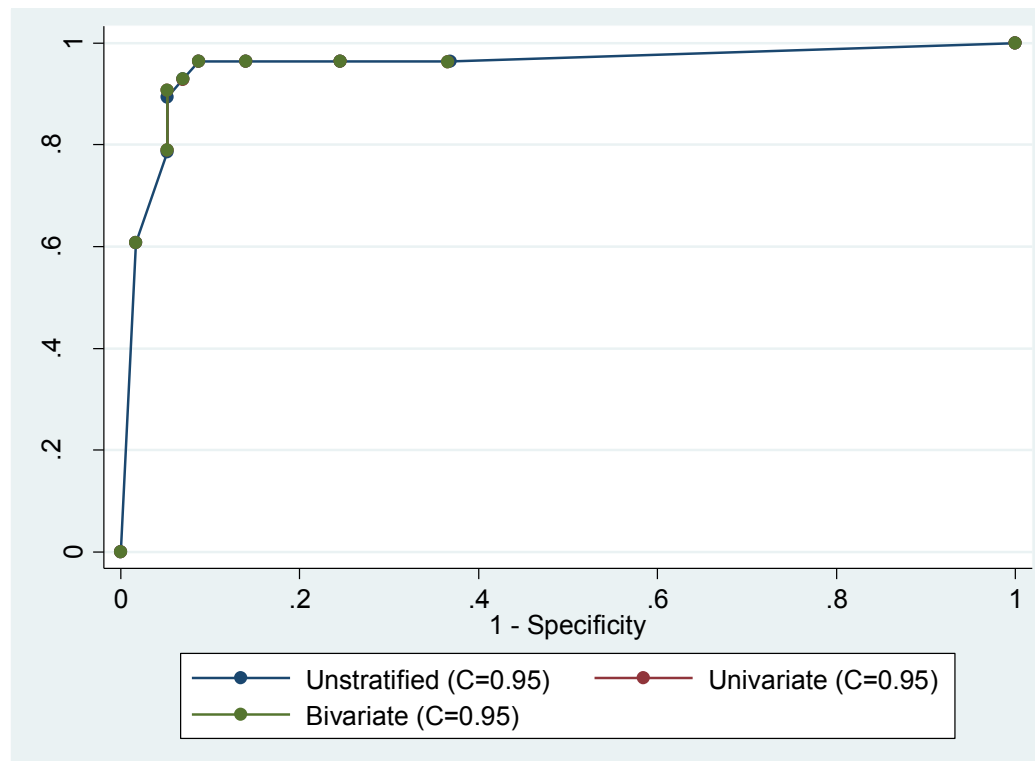
Table 4.7: Table for Youden's statistic comparing the univariate approach to the unstratified approach for 1-2 hours

1-2 hours % PTH Decrease	Youden's statistic		
	Unstratified analysis	Univariate meta-analysis	Bivariate meta-analysis
>40	47.70	47.69	47.83
>50	58.81	58.86	59.08
>60	75.90	75.90	75.90
>65	74.53	75.53	75.49
>70	81.36	81.37	81.36
>72	79.10	78.63	78.63
>80	71.24	71.80	71.92
>90	55.56	55.56	55.56

Table 4.8: Sensitivity and Specificity of PTH assay in predicting postoperative hypocalcaemia for 6 hours: unstratified, univariate and bivariate results

6 hours % PTH Decrease	Summary Sensitivity	95% CI		Summary Specificity	95% CI		T^2_{Sens}	T^2_{Spec}	Correlation
		Lower	Upper		Lower	Upper			
Unstratified analysis									
>40	96.43	82.29	99.37	63.16	50.18	74.48	-	-	-
>50	96.43	82.29	99.37	75.44	62.90	84.77	-	-	-
>60	96.43	82.29	99.37	85.96	74.68	92.71	-	-	-
>65	96.43	82.29	99.37	91.23	81.06	96.19	-	-	-
>70	92.86	77.35	98.02	92.98	83.30	97.24	-	-	-
>72	89.29	72.80	96.29	94.74	85.63	98.19	-	-	-
>80	78.57	60.46	89.79	94.74	85.63	98.19	-	-	-
>90	60.71	42.41	76.43	98.25	90.71	99.69	-	-	-
Univariate meta-analysis									
>40	96.43	78.58	99.50	63.16	50.02	74.60	0.00	0.00	-
>50	96.43	78.58	99.50	75.44	62.69	84.88	0.00	0.00	-
>60	96.43	78.58	99.50	85.97	74.36	92.82	0.00	0.00	-
>65	96.43	78.58	99.50	91.23	80.60	96.30	0.00	0.00	-
>70	92.86	75.52	98.21	92.98	82.75	97.34	0.00	0.00	-
>72	90.70	58.83	98.52	94.79	84.73	98.35	0.27	0.00	-
>80	78.89	56.57	91.47	94.75	84.84	98.31	0.02	0.00	-
>90	60.71	41.99	76.74	98.25	88.57	99.75	0.00	0.00	-
Bivariate meta-analysis									
>40	96.32	79.44	99.44	63.39	47.31	76.95	0.25	0.04	1
>50	96.43	78.58	99.50	75.44	62.69	84.88	0.00	0.00	1
>60	96.43	78.58	99.50	85.97	74.36	92.82	0.00	0.00	1
>65	96.43	78.58	99.50	91.23	80.60	96.30	0.00	0.00	0.14
>70	92.86	75.52	98.21	92.98	82.75	97.34	0.00	0.00	-1
>72	90.70	58.83	98.52	94.79	84.73	98.35	0.31	0.01	1
>80	78.89	56.57	91.47	94.75	84.84	98.31	0.04	0.00	1
>90	60.71	41.99	76.74	98.25	88.57	99.75	0.00	0.00	1

Figure 4.3: ROC curves comparing the unstratified and meta-analysis approaches for 6 hours



#Only one is visible as all three ROC curves are identical, they are superimposed on each other

Table 4.9: Table for Youden's statistic comparing the two approaches for 6 hours

6 hours % PTH Decrease	Youden's statistic		
	Unstratified analysis	Univariate meta-analysis	Bivariate meta-analysis
>40	59.59	59.71	59.71
>50	71.87	71.87	71.87
>60	82.39	82.39	82.40
>65	87.66	87.66	87.66
>70	85.84	85.84	85.84
>72	84.03	85.49	85.49
>80	73.31	73.65	73.64
>90	58.96	58.96	58.96

4.5.3 C statistic

The ROC curves for each method at each time-point of PTH measurement are presented in Figures 4.4-4.6. Due to the similar results from all three approaches, the curves were either very close or super-imposed. Now, the C statistic is compared for the unstratified and meta-analysis methods.

(i) Unstratified summary ROC curves

Figures 4.4-4.6 show the unstratified summary ROC curves for PTH measured at 0-20 minutes (C statistic=0.86 (0.80, 0.93)), 1-2 hours (C statistic=0.90 (0.85, 0.94)) and 6 hours (C statistic=0.88 (0.82, 0.95)).

Figure 4.4: 0-20 minutes unstratified ROC curve

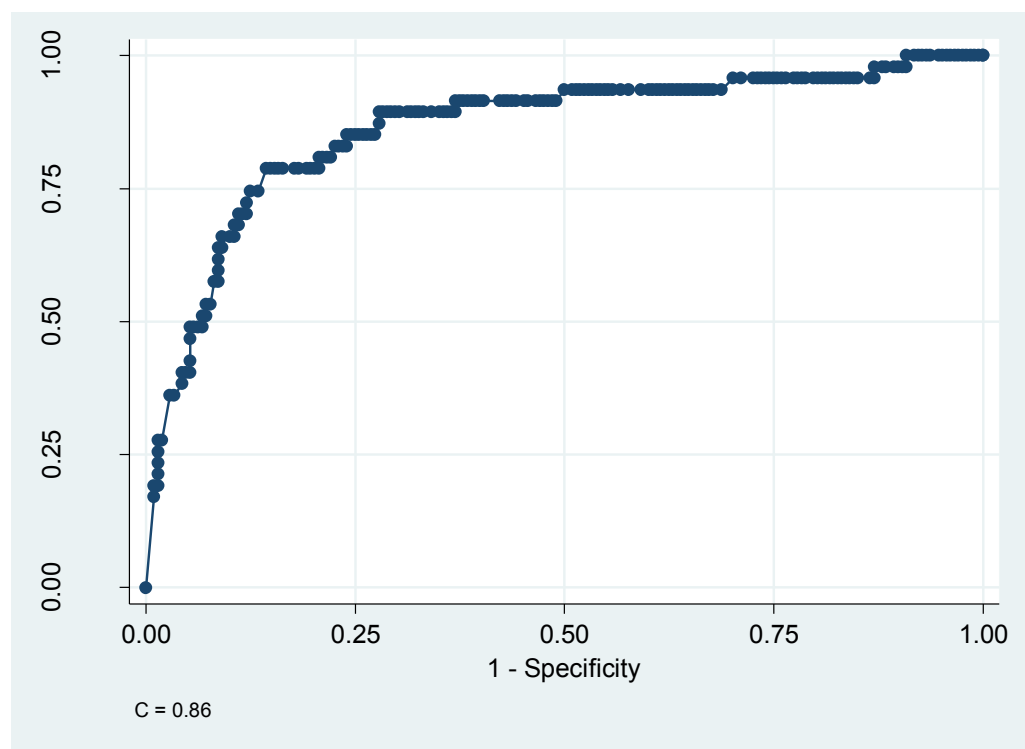


Figure 4.5: 1-2 hours unstratified ROC curve

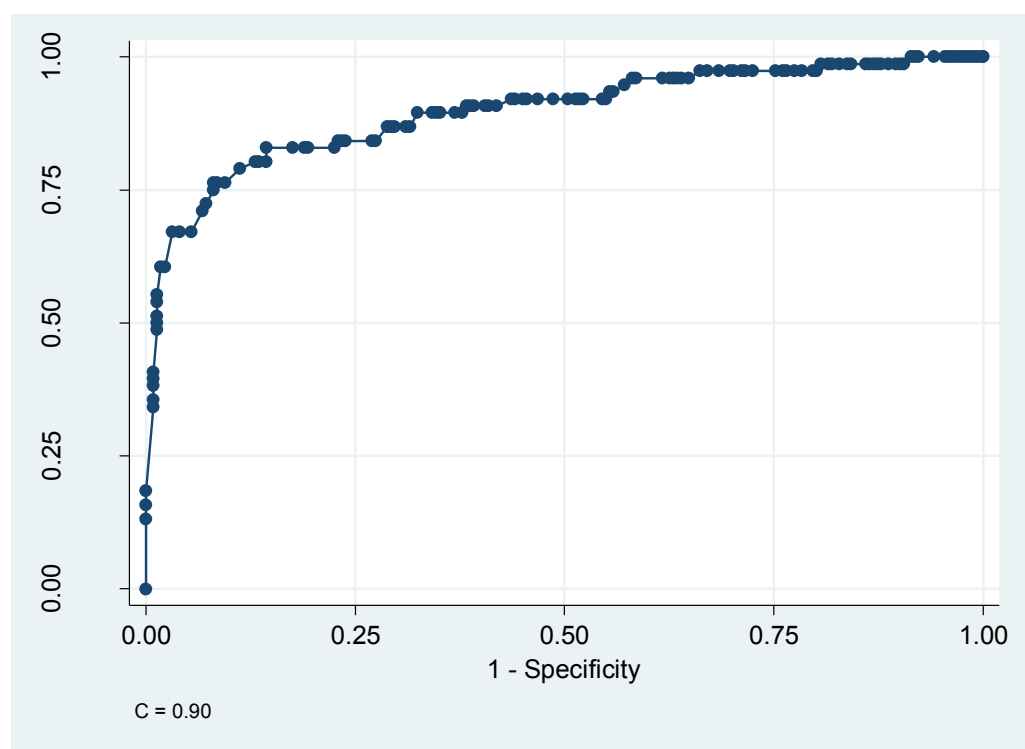
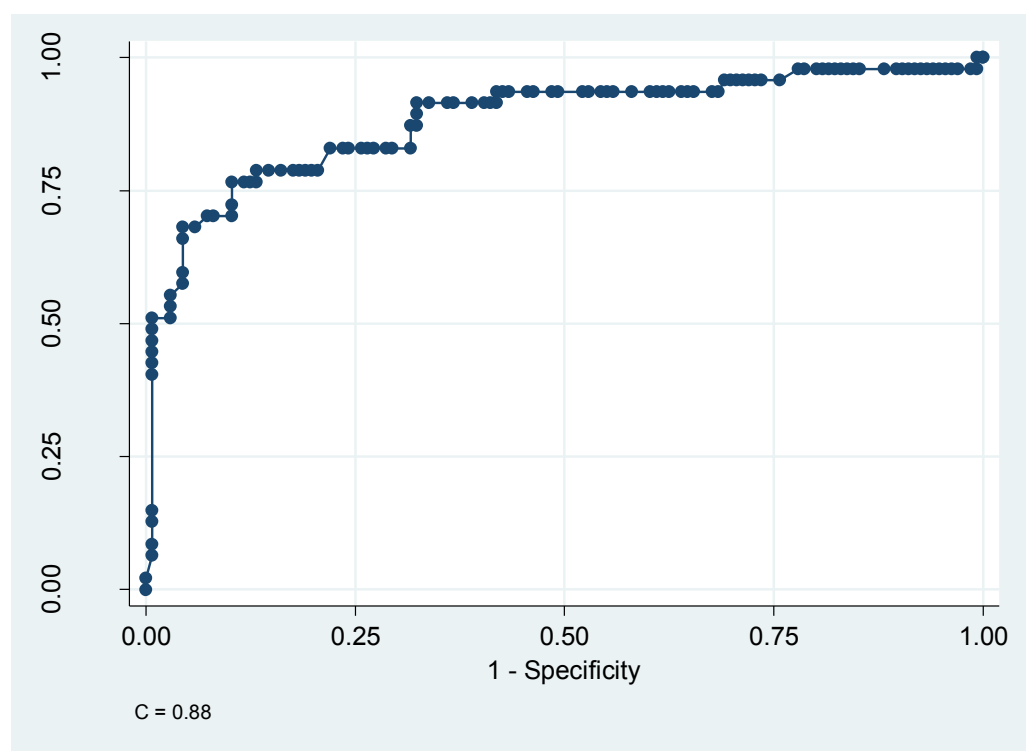


Figure 4.6: 6 hours unstratified ROC curve



(ii) Meta-analysis of C statistics

Table 4.10 shows the C statistic estimate for each study at each time-point along with its 95% confidence interval. Figures 4.7-4.9 display the ROC curves for each study at each time-point.

Table 4.10: C statistic for each study, within each time-point

Study	C statistic	95% CI		Standard Error
		Lower	Upper	
0-20 Minutes				
McLeod	0.86	0.74	0.97	0.06
Warren 02	0.70	0.35	1.00	0.18
Warren 04	0.88	0.74	1.00	0.07
Lo	0.96	0.92	1.00	0.02
Lombardi	0.92	0.84	1.00	0.04
1-2 hours				
McLeod	0.94	0.87	1.00	0.03
Warren 02	0.63	0.00	1.00	0.38
Warren 04	0.93	0.82	1.00	0.05
Lam	1.00	1.00	1.00	0.00
Lombardi	0.94	0.87	1.00	0.04
6 hours				
Lam	0.98	0.94	1.00	0.02
Lombardi	0.95	0.89	1.00	0.03

Figure 4.7: 0-20 minutes study-specific ROC curves

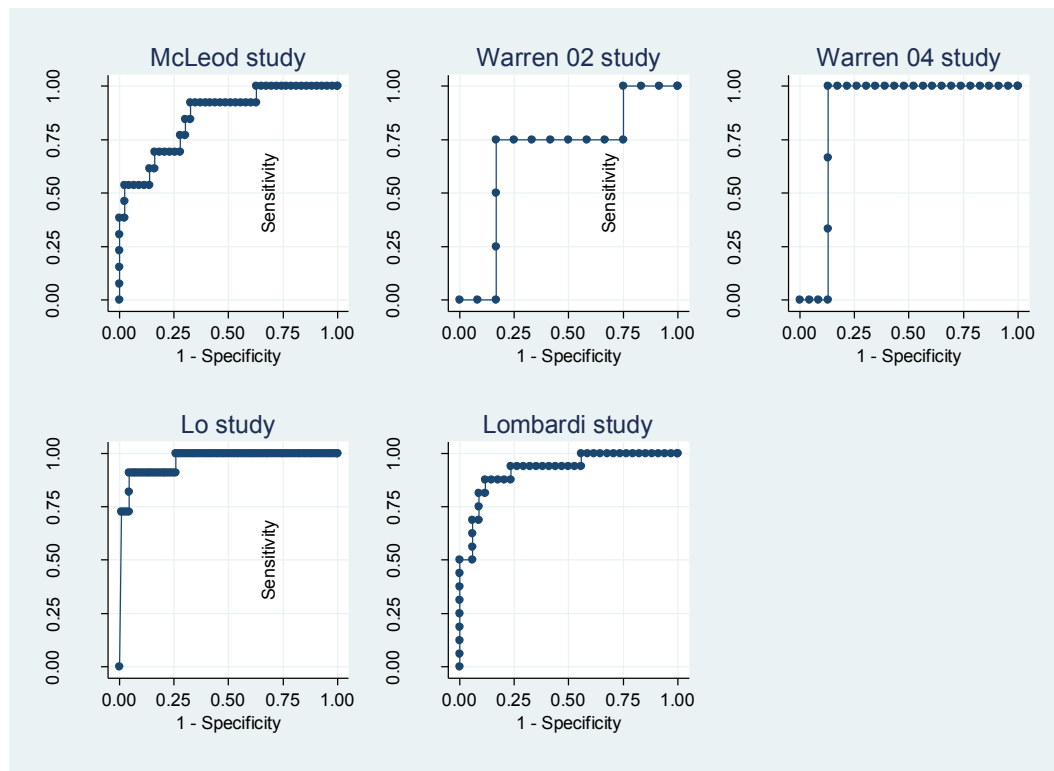


Figure 4.8: 1-2 hours study-specific ROC curves

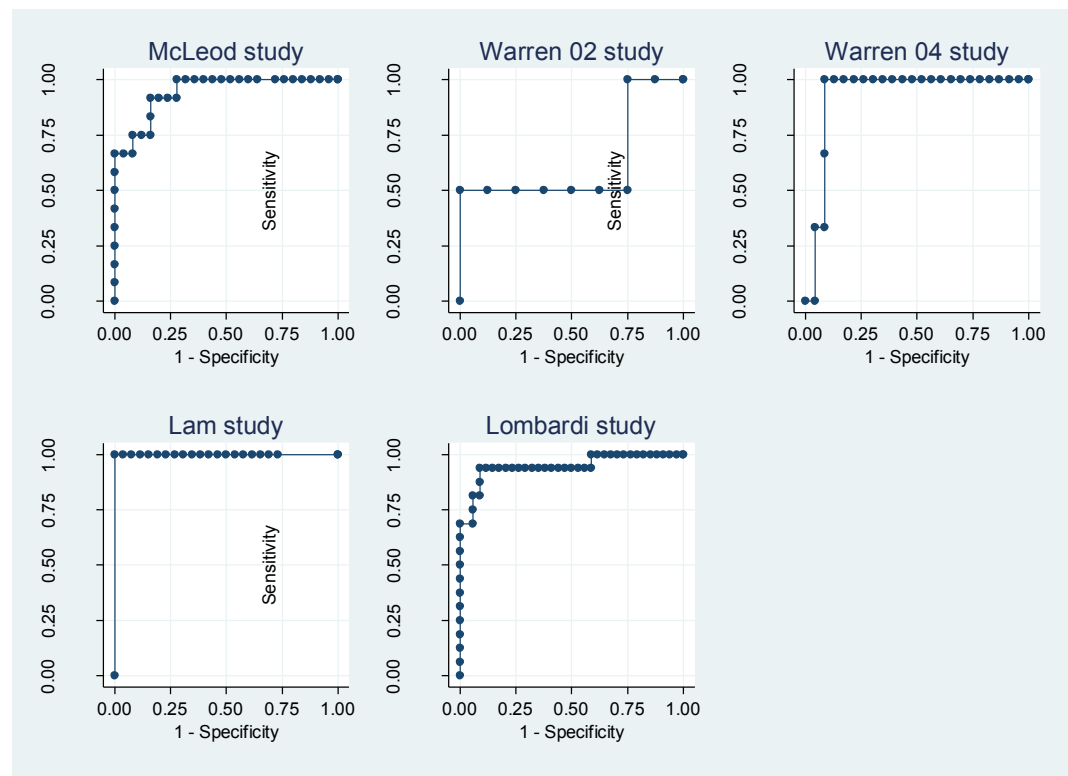
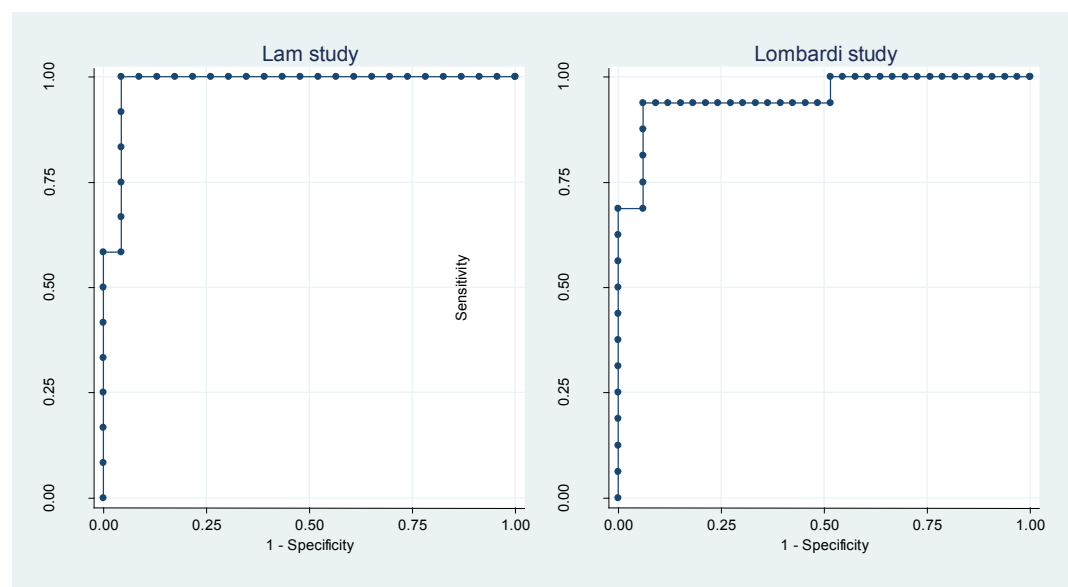


Figure 4.9: 6 hours study-specific ROC curves



A random effects meta-analysis (Equation 1.15) of the C statistic from each study was done for each time-point, and the results are summarised in Table 4.11. The summary C statistic is larger as time goes on; it is improving its accuracy in discriminating between those patients who are hypocalcaemic and those whom are not. This is expected, as the later times are closer to when the true diagnosis of hypocalcaemia is made. Although there is heterogeneity in the C statistic across studies, the 95% prediction interval suggests it will be above 0.77 in most applications, and so will always be at least moderately high. Crucially, this finding is hidden when just looking at the C statistic from the unstratified approach or the C statistic for the summary ROC curves from the univariate or bivariate meta-analysis, as heterogeneity in the C statistic is not directly modelled.

Table 4.11: C statistic random-effects meta-analysis

Time point	No of studies	Summary C statistic	95% CI		95% Prediction Interval		Tau-squared
			Lower	Upper	Lower	Upper	
0-20 min	5	0.92	0.86	0.97	0.77	1.00	<0.001
1-2 hour	5	0.94	0.89	0.99	0.84	1.00	<0.001
6 hour	2	0.97	0.94	1.00	#	#	<0.001

#Not able to estimate as <3 studies

Table 4.12: C statistic meta-analysis across all studies

Approaches	Unstratified	95% CI		Meta-analysis of	95% CI	
Time point	C statistic	Lower	Upper	C statistic	Lower	Upper
0-20 min	0.86	0.80	0.95	0.92	0.86	0.97
1-2 hour	0.90	0.85	0.94	0.94	0.89	0.99
6 hour	0.88	0.82	0.95	0.97	0.94	1.00

Table 4.12, presents the C statistics calculated in this chapter, the first-third columns refer to the unstratified results and the fourth-sixth columns refer to the values in Table 4.11. Both the approaches give similar C statistic for 1-2 hours and 6 hours, but 0-20

minutes is different with the meta-analysis value being higher 0.92 compared to 0.84. This is mainly because specificity is often estimated lower in the unstratified analysis compared to meta-analysis results, between thresholds 60 to 72, which in turn reduces the C statistic. Note: that that observed C statistics are larger from the meta-analysis approach compared to the unstratified approach, even though the 95% confidence intervals overlap considerably.

4.6 Discussion

In the systematic review of chapter 3 a major finding was that, when modelling risk prediction using IPD from multiple studies, most studies ignored clustering. To investigate this issue, and the potential consequences on the findings, in this chapter the unstratified analysis performed by Noordzij et al. [94] was replicated and compared to univariate and bivariate meta-analysis results that account for clustering. The focus has been on a single predictor and its discriminatory ability in terms of t-test results and, more importantly, the summary sensitivity, specificity and C statistics.

Methodology findings

The main methodological message from this work is that discrimination results are often similar for the unstratified and meta-analysis approaches, however important differences can arise especially when there is heterogeneity. This was most evident in the PTH results at 0-20 minutes. At a 90% PTH decrease, the summary sensitivity is 20.27% in univariate meta-analysis and 34.04% in the unstratified analysis. Even when summary pooled estimates are very similar, the unstratified approach gives too narrow confidence intervals. Further a novel finding is that a meta-analysis of C statistics can

obtain a different summary C statistic than the area under the summary ROC curve from either an unstratified, univariate or bivariate meta-analysis. For example, for PTH at 0-20 minutes, the summary C statistic when using meta-analysis was 0.92, but 0.84 in the unstratified analysis. Further, correlation is poorly estimated in most meta-analyses when there is little heterogeneity and the same threshold is used in each study; thus the bivariate method used by the Cochrane Collaboration for pooling test accuracy reviews may be unnecessary, and could often be replaced with univariate meta-analysis.

Clinical findings

The ROC curves and tables for specificity and sensitivity for 1-2 hours and 6 hours suggest these two time points were very good in terms of the PTH assay being able to accurately discriminate between patients who get hypocalcaemia and those who do not, compared to the 0-20 minute's time point which was less accurate. This is backed -up by a meta-analysis of the C statistic from each study, which found that the C statistic improves as time goes on (0-20 minutes: 0.92, 1-2 hours: 0.94 and 6 hours: 0.97) and the 95% CI gets narrower (see Table 4.11). The 95% prediction interval provided in Table 4.11 also suggests that the C statistic will be above 0.77 in most applications, and so will always be high as that is the lower bound for 0-20 minutes. Further, at thresholds around 65%, when PTH is measured 1-6 hours post-surgery, there is very little heterogeneity in sensitivity and specificity performance. Thus, the meta-analysis findings would appear to generally agree with the unstratified findings of Noordzij et al. [94]: "Patients identified as low risk for hypocalcemia could be discharged sooner. Conversely, patients identified as high risk for hypocalcemia developing could be treated earlier, potentially shortening the duration of their hypocalcemic symptoms and

hospitalization.” However, this will also be evaluated in terms of calibration in Chapter 5.

4.6.1 Limitations

The limitations in this chapter were regarding the data used. Not all studies provided pre-operative PTH levels so three studies were not used in the analysis as PTH % reduction was examined from the time-point compared to the preoperative level. No other variables or factors were provided as part of the dataset, so only PTH was available, thus rendering it not possible to assess the addition of possibly important factors such as age, sex, or diet.

A further limitation is that there is only one example to base methodology conclusions on; in particular, this dataset has very little to no heterogeneity and differences between unstratified and meta-analysis approaches are likely to be more dramatic in other datasets that contain more heterogeneity across studies.

4.6.2 What this chapter adds?

This chapter has shown that PTH values measured between 1-6 hours are a good discriminator, and using the average sensitivity and specificity values from meta-analysis is fine in this situation, as there is little to no heterogeneity present in these values when going from study to study for most thresholds. But there are papers that highlight the need to account for clustering in IPD meta-analysis [102]. Table 4.14 summarises what the problem is, what this study adds and what is required next.

Table 4.13: What Chapter four adds?

What is known / what is the problem?
<ul style="list-style-type: none"> • Investigated the issue of clustering by obtaining data from an article from the systematic review in Chapter 3 • The original analysis was performed by Noordzij and disregarded clustering • Heterogeneity was not assessed or considered within the IPD from multiple studies
What this study adds?
<ul style="list-style-type: none"> • This compares the original unstratified values to a meta-analysis version of the same values for sensitivity, specificity and C statistic and how clustering affects these values • Not much difference was found between these values due to almost no heterogeneity in the estimates • A key contribution is this study has found there is no heterogeneity in the data as this is very important to know, otherwise it would be very difficult to generalise these results across populations (important for a predictive test) • On average discrimination stays the same; at the key thresholds there is no heterogeneity which shows the values are consistent across settings • 65% found to be the optimal threshold using Youden's J statistic, backing up the original conclusions [94], assuming sensitivity and specificity are equally important • Univariate and bivariate meta-analysis results were often similar to unstratified results for sensitivity and specificity with a few occasional discrepancies • Bivariate estimates of heterogeneity are greater than univariate estimates of heterogeneity for sensitivity and specificity; important as more heterogeneity leads to wider confidence intervals for summary results • But correlation is poorly estimated in bivariate results • Important to use a meta-analysis approach to account for clustering • Heterogeneity in sensitivity and specificity needs to be quantified: a predictive test is preferred if it is consistently good in terms of discrimination across populations • Prediction intervals allow the computation of the range of potential C, sensitivity and specificity values across populations • Novel to combine C statistics in this manner, by assessing the values in each dataset and meta-analysing; and not using the overall value • This analysis has shown that by using a meta-analysis approach to summarise values such as C statistic are better (higher, which indicate better discriminative ability) than using the unstratified approach, as shown in Table 4.12 • An important finding, as it is common in literature to collect IPD and then use an unstratified approach for analysis, as found in Chapter 3
What is needed next?
<ul style="list-style-type: none"> • This analysis has shown that the original analysis indicated PTH is a good predictive factor across all studies on average • But its ability in making individualised predictions in each study was not assessed • In Chapter 5, this important extension is assessed for the calibration of this predictive test

CHAPTER 5: UNSTRATIFIED VERSUS META-ANALYSIS APPROACHES FOR EVALUATING PREDICTIVE TEST ACCURACY USING IPD:

PART II – CALIBRATION AND CHOICES OF THRESHOLD

5.1 Introduction and objectives

Sensitivity, specificity, ROC curves and C Statistics mainly inform the discriminate ability of a test for distinguishing between diseased and non-diseased. However, they do not fully reveal whether a test will be useful for making accurate predictions; in other words, whether the predicted probability of having an outcome is well calibrated for individuals. For a predictive test, calibration can be investigated by obtaining the observed number of events (O) and comparing it to the expected number of events (E) from the test, by using the E/O ratio (Expected/Observed). This was introduced in section 1.3.2.3. An assessment of calibration is therefore required to build on the examination of discrimination in Chapter 4. In Chapter 4, it was found that sensitivity and specificity have very little to no heterogeneity. Where heterogeneity exists, this can be accounted for by using random-effects models that allow variability across studies [153, 154]. But, the aim of a meta-analysis is to estimate the average performance of a test across all the populations. The objective is to translate these *average* meta-analysis

results about PTH into clinical practice. However, calibration also depends on prevalence of disease outcome, which is needed to derive PPV and NPV, and thus the expected numbers with and without the outcome. If the outcome prevalence is very different in a particular population in comparison to the average outcome prevalence across all populations, the post-test probabilities (PPV & NPV) may be inaccurate. This chapter consider methods to formally evaluate this issue. For example, one method considered is an internal-external cross-validation (IECV) approach for examining how to use existing meta-analysis results to derive PPV and NPV for use in particular clinical populations. The IECV framework provides the opportunity to examine the calibration performance of approaches (or statistical equations) that will tailor (or predict) a test's meta-analysis results for use in clinical practice. Application is made to the PTH test introduced in Chapter 4, and thus the IECV method may help to examine the calibration of PPV and NPV results for the PTH test for use in clinical practice.

In Chapter 3, in terms of model validation, it was found that rarely is external validation performed, with internal validation being preferred. A few papers used Royston's IECV approach [101]. Recall that this is a novel approach that involves removing one study from the development phase of the model, then fitting the model on the remaining data, taking the average parameter estimate values and testing performance in the excluded study. The final step involves repeating this process by rotating the omitted study and assessing the validation in all the possible scenarios. The IECV method will be used in the novel setting of test accuracy here. An approach to determine the 'best threshold'

will also be investigated, which will build on the unstratified highest sum of sensitivity and specificity method used by Noordzij et al. [94].

5.2 Statistical methods for summarising calibration

To translate meta-analysis results to an individualised probability, one can combine the sensitivity and specificity pooled results from Chapter 4 (Equations 4.2 and 4.3) with the prevalence of hypocalcaemia, using the formula given in Equations 5.1 and 5.2 (given below), to give PPV and NPV.

The issue is, however, what prevalence to use? Noordzij et al [94] used the prevalence in the overall dataset which is 0.20; however they do not consider heterogeneity, as prevalence may vary from study to study, across clinical settings. Thus, heterogeneity is assessed and different options for prevalence are considered in this chapter. The IECV approach is then introduced as a way to evaluate the approaches further.

5.2.1 Approaches for assessing calibration

At each threshold, the pooled PPV and NPV can be estimated by substituting suitable values for the sensitivity, specificity and prevalence within Equations 5.1 and 5.2. There are six approaches that are considered and summarised as follows. Note, we do not consider the summary bivariate meta-analysis results from here on, due to their similarity with the univariate meta-analysis results and the difficulty estimating the between-study correlation.

5.2.1.1 Approach 1: Unstratified

The unstratified values of sensitivity, specificity and prevalence are used to calculate PPV and NPV from Equations 5.1 and 5.2. In other words, PPV and NPV are calculated by taking the prevalence of hypocalcaemia in the unstratified data, and combining with the unstratified sensitivity and specificity, in the following equations:

$$PPV = \frac{(\text{sensitivity}) * (\text{prevalence})}{(\text{sensitivity}) * (\text{prevalence}) + (1 - \text{specificity}) * (1 - \text{prevalence})}$$

= predicted probability of developing hypocalcaemia for a patient who is test positive

Also, from a 2 by 2 table (Table 4.2):

$$PPV = \frac{a}{a + b}$$

(Equation 5.1)

$$NPV = \frac{(\text{specificity}) * (1 - \text{prevalence})}{(\text{specificity}) * (1 - \text{prevalence}) + (1 - \text{sensitivity}) * (\text{prevalence})}$$

= predicted probability of not developing hypocalcaemia for a patient who is test
negative

Also, from a 2 by 2 table (Table 4.2):

$$NPV = \frac{d}{c + d}$$

(Equation 5.2)

These predicted probabilities from (5.1) and (5.2) can be examined with the observed outcome risk using the Expected/Observed (E/O) ratio. For example for test positive patients the expected number of hypocalcaemia cases is:

Expected = (PPV) * (Number of patients testing positive in the PTH test)(Equation 5.3)
And for test negative patients it is:

$$\begin{aligned}\text{Expected} &= (1 - \text{NPV}) * (\text{Number of patients testing negative in the PTH test}) \\ &= E_2\end{aligned}$$

(Equation 5.4)

And therefore total expected $E = E_1 + E_2$

(Equation 5.5)

The E/O (Expected/Observed) ratio can then be easily obtained, where E refers to the total expected number of events or patients as predicted by the model and O refers to the total observed number of events or patients. To work out the 95% confidence interval for the E/O ratio, this formula was used, where $p = \text{Observed}/n(\text{total})$ and is given approximately by [155]:

$$95\% \text{ CI} = \frac{E}{O} * e^{\left(\pm 1.96 * \sqrt{\frac{1-p}{np}}\right)}$$

(Equation 5.6)

Where $\sqrt{\frac{1-p}{np}}$ is the standard error of $\ln(E/O)$.

5.2.1.2 Approach 2: Univariate summary sensitivity and specificity, and univariate summary prevalence

The summary meta-analysis values for sensitivity, specificity and prevalence (see Equation 5.7) are used to calculate PPV and NPV, by plugging their values into Equations 5.1 and 5.2. In this chapter only univariate meta-analysis results are considered, due to their similarity with the bivariate results (see Chapter 4) for

sensitivity and specificity. To obtain a meta-analysis summary for the prevalence, Equation 5.7 can be performed, using a binomial within-study distribution, to compute its average value, the heterogeneity in it, and a 95% prediction interval (Equation 5.8) for the prevalence in a single study.

$$\begin{aligned} r_{Outcome_i} &\sim \text{Bin}(Prev_i, N_{Everyone_i}) \\ \text{Logit}(Prev_i) &\sim N(\theta_{Prev}, \tau_{Prev}^2) \end{aligned} \quad (\text{Equation 5.7})$$

Where $r_{Outcome_i}$ is the number of patients with hypocalcemia, $Prev_i$ is the prevalence, $N_{Everyone_i}$ represents the total sample in the study i , θ_{Prev} is the average logit-prevalence across studies, τ_{Prev}^2 is the between-study heterogeneity in the logit-prevalence and the index i represents the study i in the meta-analysis.

The 95% prediction interval for the prevalence is approximately:

$$\hat{\mu} - t_{k-2} \sqrt{\tau^2 + \text{SE}(\hat{\mu})^2}, \quad \hat{\mu} + t_{k-2} \sqrt{\tau^2 + \text{SE}(\hat{\mu})^2} \quad (\text{Equation 5.8})$$

Where $\hat{\mu}$ is the estimate of the summary prevalence across studies from the random-effects meta-analysis, $\text{SE}(\hat{\mu})$ is the standard error of $\hat{\mu}$, τ is the estimate of between-study standard deviation, t_{k-2} is the $100(1 - \alpha/2)$ percentile of the t-distribution with $k-2$ degrees of freedom, where k is the number of studies in the meta-analysis and α is usually chosen as 0.05, to give a 5% significance level and thus 95% prediction interval [32].

Once PPV and NPV are derived from the meta-analysis, then E and O can be compared in each study using Equation 5.5. A random-effects meta-analysis of $\ln(E/O)$ can then be performed using Equation 1.15.

5.2.1.3 Approach 3: Univariate summary sensitivity and specificity, lower predicted prevalence

The summary univariate meta-analysis values for sensitivity and specificity (see Equations 4.5-4.6) are combined with the lower 95% prediction value for prevalence (see Equation 5.8) to calculate PPV and NPV using (5.1) and (5.2).

5.2.1.4 Approach 4: Univariate summary sensitivity and specificity, upper predicted prevalence

The summary univariate values for sensitivity and specificity (see Equations 4.5-4.6) along with the upper 95% prediction value for prevalence (see Equations 5.8) are used to calculate PPV and NPV using (5.1) and (5.2).

These approaches (3 & 4) have been used to simply show the extreme variability in deriving PPV and NPV results when using the lower and upper prevalence's as given by the 95% prediction interval. This illustrates what happens when we take a value of prevalence that is quite different to the summary (average) prevalence across studies, and that – when applying our test accuracy results in practice - we need to be sure that the chosen prevalence is pertinent to the population for application.

5.2.1.5 Approach 5: Univariate summary sensitivity and specificity, study-specific prevalence

The summary univariate values for sensitivity and specificity (see Equations 4.5-4.6) along with the observed prevalence from an individual study are used to calculate PPV and NPV via (5.1) and (5.2). The observed prevalence is study-specific, as it is for those patients within that study setting.

5.2.1.6 Approach 6: Internal-external validation

Approach 6 uses the internal-external cross-validation (IECV) approach to examine different choices of the prevalence when tailoring test accuracy results about PPV and NPV to the clinical populations of interest. Indeed, all of approaches (1) to (5) can rather be embedded in this framework, where a study is omitted so that the performance of meta-analysis PPV and NPV results (derived from other studies) is then evaluated in it. However, for brevity only approach 5 is examined using IECV (that is, only the use of the study-specific prevalence combined with summary meta-analysis results for sensitivity and specificity to derive PPV and NPV is evaluated using IECV). To generate and validate NPV and PPV within the IECV framework, the following nine steps are used:

- i. Select a study l to be excluded
- ii. In the remaining $k-1$ studies, fit a univariate meta-analysis model (Equations 4.5-4.6) and obtain summary values for sensitivity and specificity
- iii. Choose a prevalence for the excluded study; here the observed prevalence in the excluded study is used (akin to approach 5)

- iv. Calculate PPV and NPV using the summary sensitivity and specificity values from step ii and the prevalence from step iii, by plugging them into equations (5.1) and (5.2)
- v. In the excluded study l , calculate total observed (O) and total expected (E) outcomes, using equations 5.3-5.5
- vi. In the excluded study, calculate (E/O), $\ln(E/O)$ and the standard error of $\ln(E/O)$ as stated in equation 5.6
- vii. Repeat steps (i) to (vi) for each excluded study, giving a set of k values for $\ln(E/O)$ and its standard error
- viii. Perform a random-effects meta-analysis of the $\ln(E/O)_l$ values and then transform the results back to (E/O)
- ix. Finally summarise the calibration across all the studies by displaying the meta-analysis results of the summary E/O value, its 95% confidence interval and a 95% prediction interval in a forest plot

IECV is considered relevant to Chapter 5 (calibration) but not Chapter 4 (discrimination), because discrimination is independent of the prevalence (intercept) term, as it is based on just the predictors (here, the single predictor of PTH). There is not a meta-analysis equation to predict the *exact* C statistic in a new study (only the range of possible C statistic values from a prediction interval), and thus IECV is not relevant to examine predictive performance (this is akin to not using IECV in a meta-analysis of trials that aim to summarise a treatment effect; the meta-analyst may derive a

prediction interval for the range of possible treatment effects in a new population, but do not usually want to predict the exact new treatment effect).

However, calibration depends on both the prevalence (intercept) and the predictors, and the total equation is used to provide risk predictions. It is then of crucial interest for these risk predictions to calibrate well in new data, independent to the data used to estimate the predictor effects (and ideally the intercept). Therefore, IECV is more important for the calibration, to check whether predictions from the model agree with observed risks in new data, and hence why IECV was used in this Chapter 5 but not 4. Further, the choice of intercept term in the prediction equation is optional, and the IECV allows the user to examine which is the best intercept to take (summary prevalence, or tailored prevalence to new population) to improve the accuracy of predictions

5.3 Results I – Calibration of PPV and NPV

The six IPD studies available are summarised in Table 5.1. This table was described already in detail in Chapter 4 but is presented here again for ease. The prevalence of hypocalcaemia varies across studies, from 0.11 to 0.31, and the unstratified prevalence in the pooled IPD is 0.20.

A random-effects meta-analysis of the prevalence estimates, using Equation 5.6, gave an average prevalence of 0.22, similar to the prevalence when not accounting for clustering in the data, but a 95% prediction interval of 0.09 to 0.42 reveals the heterogeneity and that across settings; prevalence could be between 0.09 and 0.42. As prevalence can vary considerably across settings, this could affect the calibration

performance of the PTH test when applied to individual settings by using the average. Note that all six studies have been used in this chapter. But not all six provide data at each time point i.e. 0-20 minutes (McLeod, Warren 02, Warren 04, Lo and Lombardi), 1-2 hours (McLeod, Warren 02, Warren 04, Lam and Lombardi) and 6 hours (Lam and Lombardi).

Table 5.1: Summary of the 6 IPD studies

Study	No of patients in reported article	No of events in reported article	Outcome Prevalence in reported article	Prospective?	Definition of hypocalcaemia (equivalent cCa (mg/dL))	IPD provided: no. of patients (events)	Outcome Prevalence in IPD	Times at which PTH was measured
Lam 2003 [141]	40	12	0.30	Yes	<7.6	39 (12)	0.31	Pre-op, 1 and 6 hour post-op
Lo 2002 [142]	155	32	0.21	Yes	<7.2	100 (11)	0.11	Pre-op and intra-op
Lombardi 2004 [143]	53	16	0.30	Yes	<8.0	52 (16)	0.31	Pre-op, intra-op, 2, 4, 6, 24 and 48 hour post-op
McLeod 2006 [144]	60	15	0.25	Yes	<8.0	60 (14)	0.23	Pre-op, intra-op and 1 hour post-op
Warren 2002 [145]	53	5	0.09	Yes	<8.0	22 (4)	0.18	Pre-op, Intra-op and 1 hour post-op
Warren 2004 [146]	27	3	0.11	Yes	<8.0	27 (3)	0.11	Pre-op, intra-op and 1 hour post-op
Overall	388	83	0.21			300(60)	0.20	

5.3.1 Comparison of PPV and NPV estimates

The NPV & PPV are now compared across the first five approaches in this section.

Approach 6, which is the internal-external validation method, will be assessed in section 5.3.4.

Table 5.2: PPV and NPV for 0-20 minutes

0-20 minutes % PTH Decrease	Prevalence	Sensitivity	Specificity	PPV	NPV
Approach (1)	Unstratified	Unstratified	Unstratified		
>40	0.21	93.62	44.06	31.00	96.26
>50	0.21	93.62	56.44	36.59	97.05
>60	0.21	85.11	70.79	43.89	94.66
>65	0.21	80.85	74.26	45.75	93.53
>70	0.21	80.85	75.74	47.22	93.64
>72	0.21	80.85	76.73	48.26	93.72
>80	0.21	68.09	85.64	56.00	90.91
>90	0.21	34.04	95.54	67.20	84.36
Approach (2)	Univariate	Univariate	Univariate		
>40	0.22	93.62	46.75	32.57	96.39
>50	0.22	93.62	56.44	37.12	96.99
>60	0.22	85.11	75.58	48.91	94.87
>65	0.22	80.85	79.05	51.46	93.76
>70	0.22	80.85	79.78	52.35	93.81
>72	0.22	80.85	81.55	54.62	93.94
>80	0.22	70.21	85.15	56.50	91.23
>90	0.22	20.27	96.17	59.25	81.45
Approach (3)	Lower 95% PI	Univariate	Univariate		
>40	0.09	93.62	46.75	15.48	98.60
>50	0.09	93.62	56.44	18.30	98.84
>60	0.09	85.11	75.58	26.64	97.99
>65	0.09	80.85	79.05	28.68	97.54
>70	0.09	80.85	79.78	29.41	97.56
>72	0.09	80.85	81.55	31.34	97.61
>80	0.09	70.21	85.15	33.00	96.48
>90	0.09	20.27	96.17	35.54	92.05
Approach (4)	Upper 95% PI	Univariate	Univariate		
>40	0.42	93.62	46.75	56.02	91.00
>50	0.42	93.62	56.44	60.89	92.43
>60	0.42	85.11	75.58	71.63	87.51
>65	0.42	80.85	79.05	73.65	85.07
>70	0.42	80.85	79.78	74.34	85.19
>72	0.42	80.85	81.55	76.04	85.46
>80	0.42	70.21	85.15	77.40	79.78
>90	0.42	20.27	96.17	79.31	62.48
Approach (5)	Studies own prevalence	Univariate	Univariate		
McLeod					
>40	0.23	93.62	46.75	34.86	96.01
>50	0.23	93.62	56.44	39.54	96.67
>60	0.23	85.11	75.58	51.47	94.34
>65	0.23	80.85	79.05	54.01	93.13
>70	0.23	80.85	79.78	54.89	93.19
>72	0.23	80.85	81.55	57.15	93.33
>80	0.23	70.21	85.15	59.00	90.38
>90	0.23	20.27	96.17	61.70	79.85
Warren 02					
>40	0.18	93.62	46.75	28.09	97.06
>50	0.18	93.62	56.44	32.32	97.55
>60	0.18	85.11	75.58	43.65	95.81
>65	0.18	80.85	79.05	46.17	94.89
>70	0.18	80.85	79.78	47.05	94.94
>72	0.18	80.85	81.55	49.34	95.04
>80	0.18	70.21	85.15	51.24	92.79
>90	0.18	20.27	96.17	54.05	84.44
Warren 04					
>40	0.11	93.62	46.75	18.02	98.32
>50	0.11	93.62	56.44	21.17	98.61
>60	0.11	85.11	75.58	30.34	97.60

0-20 minutes % PTH Decrease	Prevalence	Sensitivity	Specificity	PPV	NPV
>65	0.11	80.85	79.05	32.54	97.06
>70	0.11	80.85	79.78	33.32	97.09
>72	0.11	80.85	81.55	35.39	97.15
>80	0.11	70.21	85.15	37.14	95.81
>90	0.11	20.27	96.17	39.81	90.61
Lo					
>40	0.11	93.62	46.75	17.85	98.34
>50	0.11	93.62	56.44	20.99	98.62
>60	0.11	85.11	75.58	30.11	97.62
>65	0.11	80.85	79.05	32.29	97.09
>70	0.11	80.85	79.78	33.07	97.12
>72	0.11	80.85	81.55	35.13	97.18
>80	0.11	70.21	85.15	36.88	95.86
>90	0.11	20.27	96.17	39.54	90.71
Lombardi					
>40	0.31	93.62	46.75	43.86	94.28
>50	0.31	93.62	56.44	48.86	95.22
>60	0.31	85.11	75.58	60.77	91.95
>65	0.31	80.85	79.05	63.17	90.28
>70	0.31	80.85	79.78	63.99	90.36
>72	0.31	80.85	81.55	66.07	90.55
>80	0.31	70.21	85.15	67.76	86.54
>90	0.31	20.27	96.17	70.17	73.07

Table 5.2 shows for PTH at 0-20 minutes that the PPV can be very different when using the summary results from a univariate meta-analysis (approach 2) compared to the unstratified results (approach 1). For example, at a threshold of 72% the PPV is 48% using the unstratified value and 55% using the meta-analysis values, but at threshold 90% it is lower, 59% (approach 2) compared to 67% (approach 1). Differences also occur for NPV; e.g. at a 90% threshold, unstratified (approach 1) and meta-analysis (approach 2) approaches give a NPV of 84% and 81% respectively.

When the lower and upper bounds of the prevalence's prediction interval are used (approaches 3 and 4), the NPV and PPV considerably change as expected. This reveals how important the choice of prevalence is. For example, at a 72% threshold the PPV is 55% based on the meta-analysis average prevalence, 31% based on the lower prediction bound and 76% based on the upper prediction bound. Tables 5.3 and 5.4 reveal similar findings for 1-2 hours and 6 hours.

Table 5.3: PPV and NPV for 1-2 hours

1-2 hours % PTH Decrease	Prevalence	Sensitivity	Specificity	PPV	NPV
Approach (1)	Unstratified	Unstratified	Unstratified		
>40	0.21	95.56	52.14	34.89	97.77
>50	0.21	95.56	63.25	41.11	98.15
>60	0.21	95.56	80.34	56.61	98.54
>65	0.21	93.33	81.20	57.13	97.84
>70	0.21	93.33	88.03	67.67	98.01
>72	0.21	89.36	89.74	70.04	96.92
>80	0.21	77.78	93.46	76.15	94.00
>90	0.21	55.56	100.0	100.00	89.34
Approach (2)	Univariate	Univariate	Univariate		
>40	0.22	95.56	52.14	35.42	97.71
>50	0.22	95.56	63.31	41.71	98.11
>60	0.22	95.56	80.34	57.18	98.50
>65	0.22	93.33	82.19	59.01	97.82
>70	0.22	93.33	88.03	68.17	97.96
>72	0.22	88.89	89.74	70.42	96.71
>80	0.22	77.78	94.02	78.13	93.90
>90	0.22	55.56	100.00	100.00	89.12
Approach (3)	Lower 95% PI	Univariate	Univariate		
>40	0.09	95.56	52.14	17.22	99.12
>50	0.09	95.56	63.31	21.34	99.27
>60	0.09	95.56	80.34	33.62	99.43
>65	0.09	93.33	82.19	35.32	99.16
>70	0.09	93.33	88.03	44.82	99.22
>72	0.09	88.89	89.74	47.44	98.73
>80	0.09	77.78	94.02	57.54	97.60
>90	0.09	55.56	100.00	100.00	95.57
Approach (4)	Upper 95% PI	Univariate	Univariate		
>40	0.42	95.56	52.14	59.12	94.19
>50	0.42	95.56	63.31	65.36	95.17
>60	0.42	95.56	80.34	77.88	96.15
>65	0.42	93.33	82.19	79.15	94.45
>70	0.42	93.33	88.03	84.96	94.80
>72	0.42	88.89	89.74	86.26	91.77
>80	0.42	77.78	94.02	90.40	85.38
>90	0.42	55.56	100.00	100.00	75.65
Approach (5)	Studies own prevalence	Univariate	Univariate		
McLeod					
>40	0.23	95.56	52.14	37.36	97.52
>50	0.23	95.56	63.31	43.76	97.95
>60	0.23	95.56	80.34	59.21	98.38
>65	0.23	93.33	82.19	61.02	97.63
>70	0.23	93.33	88.03	69.96	97.79
>72	0.23	88.89	89.74	72.13	96.43
>80	0.23	77.78	94.02	79.53	93.41
>90	0.23	55.56	100.00	100.00	88.28
Warren 02					
>40	0.18	95.56	52.14	30.47	98.17
>50	0.18	95.56	63.31	36.38	98.48
>60	0.18	95.56	80.34	51.62	98.80
>65	0.18	93.33	82.19	53.50	98.25
>70	0.18	93.33	88.03	63.12	98.36
>72	0.18	88.89	89.74	65.54	97.35
>80	0.18	77.78	94.02	74.06	95.07
>90	0.18	55.56	100.00	100.00	91.11
Warren 04					
>40	0.11	95.56	52.14	19.79	98.96
>50	0.11	95.56	63.31	24.35	99.14
>60	0.11	95.56	80.34	37.53	99.32

1-2 hours % PTH Decrease	Prevalence	Sensitivity	Specificity	PPV	NPV
>65	0.11	93.33	82.19	39.31	99.01
>70	0.11	93.33	88.03	49.08	99.07
>72	0.11	88.89	89.74	51.71	98.49
>80	0.11	77.78	94.02	61.65	97.16
>90	0.11	55.56	100.00	100.00	94.79
Lam					
>40	0.31	95.56	52.14	47.29	96.32
>50	0.31	95.56	63.31	53.92	96.95
>60	0.31	95.56	80.34	68.59	97.58
>65	0.31	93.33	82.19	70.19	96.48
>70	0.31	93.33	88.03	77.79	96.71
>72	0.31	88.89	89.74	79.56	94.73
>80	0.31	77.78	94.02	85.39	90.40
>90	0.31	55.56	100.00	100.00	83.36
Lombardi					
>40	0.31	95.56	52.14	47.29	96.32
>50	0.31	95.56	63.31	53.92	96.95
>60	0.31	95.56	80.34	68.59	97.58
>65	0.31	93.33	82.19	70.19	96.48
>70	0.31	93.33	88.03	77.79	96.71
>72	0.31	88.89	89.74	79.56	94.73
>80	0.31	77.78	94.02	85.39	90.40
>90	0.31	55.56	100.00	100.00	83.36

Table 5.4: PPV and NPV for 6 hours

6hr % PTH Decrease	Prevalence	Sensitivity	Specificity	PPV	NPV
Approach (1)					
>40	0.21	96.43	63.16	41.27	98.51
>50	0.21	96.43	75.44	51.31	98.75
>60	0.21	96.43	85.96	64.83	98.90
>65	0.21	96.43	91.23	74.69	98.96
>70	0.21	92.86	92.98	78.03	97.98
>72	0.21	89.29	94.74	82.00	97.05
>80	0.21	78.57	94.74	80.04	94.28
>90	0.21	60.71	98.25	90.30	90.31
Approach (2)					
>40	0.22	96.43	63.16	41.83	98.47
>50	0.22	96.43	75.44	51.89	98.72
>60	0.22	96.43	85.97	65.38	98.87
>65	0.22	96.43	91.23	75.13	98.94
>70	0.22	92.86	92.98	78.42	97.93
>72	0.22	90.70	94.79	82.71	97.38
>80	0.22	78.89	94.75	80.50	94.23
>90	0.22	60.71	98.25	90.50	90.10
Approach (3)					
	Lower 95% PI	Univariate	Univariate		
>40	0.09	96.43	63.16	21.43	99.41
>50	0.09	96.43	75.44	29.03	99.51
>60	0.09	96.43	85.97	41.73	99.57
>65	0.09	96.43	91.23	53.39	99.59
>70	0.09	92.86	92.98	57.95	99.21
>72	0.09	90.70	94.79	64.46	98.99
>80	0.09	78.89	94.75	61.02	97.73
>90	0.09	60.71	98.25	78.33	96.00
Approach (4)					
	Upper 95% PI	Univariate	Univariate		
>40	0.42	96.43	63.16	65.47	96.07
>50	0.42	96.43	75.44	73.99	96.69

6hr % PTH Decrease	Prevalence	Sensitivity	Specificity	PPV	NPV
>60	0.42	96.43	85.97	83.27	97.08
>65	0.42	96.43	91.23	88.85	97.24
>70	0.42	92.86	92.98	90.55	94.73
>72	0.42	90.70	94.79	92.65	93.36
>80	0.42	78.89	94.75	91.59	86.10
>90	0.42	60.71	98.25	96.17	77.54
Approach (5)	Studies own prevalence	Univariate	Univariate		
Lam					
>40	0.31	96.43	63.16	53.78	97.55
>50	0.31	96.43	75.44	63.57	97.94
>60	0.31	96.43	85.97	75.34	98.19
>65	0.31	96.43	91.23	83.01	98.29
>70	0.31	92.86	92.98	85.46	96.7
>72	0.31	90.70	94.79	88.55	95.82
>80	0.31	78.89	94.75	86.98	90.99
>90	0.31	60.71	98.25	93.91	84.91
Lombardi					
>40	0.31	96.43	63.16	53.78	97.55
>50	0.31	96.43	75.44	63.57	97.94
>60	0.31	96.43	85.97	75.34	98.19
>65	0.31	96.43	91.23	83.01	98.29
>70	0.31	92.86	92.98	85.46	96.7
>72	0.31	90.70	94.79	88.55	95.82
>80	0.31	78.89	94.75	86.98	90.99
>90	0.31	60.71	98.25	93.91	84.91

Tables 5.2-5.4 also use a fifth approach which uses each study's observed prevalence along with the summary univariate meta-analysis values for sensitivity and specificity.

5.3.2 Comparison of the calibration of predicted values

When choosing PPV and NPV values, it is important to know that they calibrate well when applied to the populations of interest, and so the E/O statistic is now examined for approaches 1 to 5, at each time-point for measuring PTH.

5.3.2.1 0-20 minutes

First the E/O ratio is estimated when the IPD are regarded as coming from one big study, using the unstratified prevalence combined with unstratified sensitivity and specificity (approach 1) to derive PPV and NPV, and then rather the univariate prevalence combined with univariate sensitivity and specificity (approach 2) to derive PPV and NPV.

Table 5.5: E/O ratios for all the patients combined using the unstratified and univariate meta-analysis approaches at the 0-20 minute's time-point

Overall	No. of hypocalcaemia who tested positive			95% CI		No. of hypocalcaemia who tested negative			95% CI		Total with hypocalcaemia			95% CI	
	E	O	E/O	Lower	Upper	E	O	E/O	Lower	Upper	E	O	E/O	Lower	Upper
<i>Approach 1</i>															
>40	48.67	44	1.11	0.82	1.49	3.44	3	1.15	0.37	3.56	52.11	47	1.11	0.83	1.48
>50	48.30	44	1.10	0.82	1.48	3.45	3	1.15	0.37	3.57	51.75	47	1.10	0.83	1.47
>60	43.45	40	1.09	0.80	1.48	8.01	7	1.14	0.55	2.40	51.46	47	1.09	0.82	1.46
>65	41.18	38	1.08	0.79	1.49	10.29	9	1.14	0.59	2.20	51.46	47	1.09	0.82	1.46
>70	41.08	38	1.08	0.79	1.49	10.30	9	1.14	0.60	2.20	51.38	47	1.09	0.82	1.46
>72	41.02	38	1.08	0.79	1.48	10.30	9	1.14	0.60	2.20	51.32	47	1.09	0.82	1.45
>80	34.16	32	1.07	0.75	1.51	17.09	15	1.14	0.69	1.89	51.25	47	1.09	0.82	1.45
>90	16.80	16	1.05	0.64	1.71	35.03	31	1.13	0.79	1.61	51.83	47	1.10	0.83	1.47
<i>Approach 2</i>															
>40	51.13	44	1.16	0.86	1.56	3.32	3	1.11	0.36	3.43	54.46	47	1.16	0.87	1.54
>50	49.00	44	1.11	0.83	1.50	3.52	3	1.17	0.38	3.64	52.52	47	1.12	0.84	1.49
>60	48.42	40	1.21	0.89	1.65	7.70	7	1.10	0.52	2.31	56.12	47	1.19	0.90	1.59
>65	46.31	38	1.22	0.89	1.67	9.92	9	1.10	0.57	2.12	56.24	47	1.20	0.90	1.59
>70	45.54	38	1.20	0.87	1.65	10.03	9	1.11	0.58	2.14	55.57	47	1.18	0.89	1.57
>72	46.43	38	1.22	0.89	1.68	9.94	9	1.10	0.57	2.12	56.37	47	1.20	0.90	1.60
>80	34.47	32	1.08	0.76	1.52	16.49	15	1.10	0.66	1.82	50.95	47	1.08	0.81	1.44
>90	14.81	16	0.93	0.57	1.51	41.55	31	1.34	0.94	1.91	56.36	47	1.20	0.90	1.60

Table 5.5 shows the combined E/O for all the patients at the 0-20 minutes time point for the unstratified approach is around 1.10 for all the thresholds; thus, the PPV and NPV derived from the unstratified approach slightly over predict. The 95% CI includes 1, reflecting a wide CI due to small number of events.

Interestingly, when using the PPV and NPV values based on univariate meta-analysis, the combined E/O for all threshold points is even further away from 1, with it being around 1.2. The 95% CI again includes 1. Thus it might appear that using the unstratified approach to derive PPV and NPV is better in terms of calibration performance of the predictive test. However, assessment of E/O in the unstratified approach is naive and misleading, as we have already seen that prevalence varies across studies. So let's now check calibration in each study separately.

First, the McLeod [144] study is presented, and all 5 approaches are considered for calculating PPV and NPV, to then obtain E and O.

Table 5.6 shows the E/O ratios, and clearly when using the unstratified approach (1) to derive PPV and NPV for clinical practice, there is an under-prediction of the numbers of events in this study, for example with an E/O of 0.85 for the 65% threshold. Using approach (2) to derive PPV and NPV gives improved calibration performance of the predictive test with E/O values closer to 1, for example with an E/O ratio of 0.93 at 65%. Using PPV and NPV estimates derived from the prevalence based on the lower (approach 3) or upper (approach 4) prediction interval values, dramatically over-predicts or under-predicts the calibration performance of the predictive test

Crucially, it seems using approach (5) (which uses the observed prevalence of 0.23 in this clinical population of interest) to derive PPV and NPV helps to improve calibration performance of the predictive test. This approach derives PPV and NPV by using each study's own observed prevalence and combining that with the univariate meta-analysis values for sensitivity and specificity, which gives an the E/O statistic very close to 1 for all thresholds. For example, an E/O of 0.98 at the threshold of 65% shows that it is comparatively better at calibration performance of the predictive test (PPV and NPV) than using PPV and NPV from other approaches, and so using the study's observed prevalence helps to improve calibration. Thus, due to heterogeneity in prevalence, using an unstratified or meta-analytic summary of prevalence is inadequate for individual settings, and results are only improved by tailoring prevalence to the intended population.

Table 5.6: E/O ratios using 5 approaches for McLeod study at the 0-20 minute's time-point

Approach	No. of hypocalcaemia who tested positive			95% CI		No. of hypocalcaemia who tested negative			95% CI		Total with hypocalcaemia			95% CI	
Threshold	E	O	E/O	Lower	Upper	E	O	E/O	Lower	Upper	E	O	E/O	Lower	Upper
(1) Using the Unstratified prevalence, Unstratified sensitivity and Unstratified specificity															
>40	11.16	12	0.93	0.53	1.64	0.75	1	0.75	0.11	5.31	11.91	13	0.92	0.53	1.58
>50	11.71	12	0.98	0.55	1.72	0.71	1	0.71	0.10	5.03	12.42	13	0.96	0.55	1.64
>60	9.66	10	0.97	0.52	1.79	1.82	3	0.61	0.20	1.88	11.47	13	0.88	0.51	1.52
>65	8.69	9	0.97	0.50	1.86	2.39	4	0.60	0.22	1.59	11.09	13	0.85	0.50	1.47
>70	8.50	9	0.94	0.49	1.82	2.42	4	0.60	0.23	1.61	10.92	13	0.84	0.49	1.45
>72	8.20	9	0.91	0.47	1.75	2.45	4	0.61	0.23	1.63	10.65	13	0.82	0.48	1.41
>80	6.72	7	0.96	0.46	2.01	4.00	6	0.67	0.30	1.48	10.72	13	0.82	0.48	1.42
>90	2.69	4	0.67	0.25	1.79	8.13	9	0.90	0.47	1.74	10.82	13	0.83	0.48	1.43
(2) Using the univariate prevalence, univariate sensitivity and univariate specificity															
>40	11.73	12	0.98	0.55	1.72	0.72	1	0.72	0.10	5.13	12.45	13	0.96	0.56	1.65
>50	11.88	12	0.99	0.56	1.74	0.72	1	0.72	0.10	5.13	12.60	13	0.97	0.56	1.67
>60	10.76	10	1.08	0.58	2.00	1.74	3	0.58	0.19	1.80	12.50	13	0.96	0.56	1.66
>65	9.78	9	1.09	0.57	2.09	2.31	4	0.58	0.22	1.54	12.09	13	0.93	0.54	1.60
>70	9.42	9	1.05	0.54	2.01	2.35	4	0.59	0.22	1.57	11.78	13	0.91	0.53	1.56
>72	9.29	9	1.03	0.54	1.98	2.36	4	0.59	0.22	1.57	11.65	13	0.90	0.52	1.54
>80	6.78	7	0.97	0.46	2.03	3.86	6	0.64	0.29	1.43	10.64	13	0.82	0.48	1.41
>90	2.37	4	0.59	0.22	1.58	9.65	9	1.07	0.56	2.06	12.02	13	0.92	0.54	1.59
(3) Using the lower 95% prediction interval bound prevalence, univariate sensitivity and univariate specificity															
>40	5.57	12	0.46	0.26	0.82	0.28	1	0.28	0.04	1.99	5.85	13	0.45	0.26	0.78
>50	5.86	12	0.49	0.28	0.86	0.28	1	0.28	0.04	1.98	6.13	13	0.47	0.27	0.81
>60	5.86	10	0.59	0.32	1.09	0.68	3	0.23	0.07	0.71	6.54	13	0.50	0.29	0.87
>65	5.45	9	0.61	0.32	1.16	0.91	4	0.23	0.09	0.61	6.36	13	0.49	0.28	0.84
>70	5.29	9	0.59	0.31	1.13	0.93	4	0.23	0.09	0.62	6.22	13	0.48	0.28	0.82
>72	5.33	9	0.59	0.31	1.14	0.93	4	0.23	0.09	0.62	6.26	13	0.48	0.28	0.83
>80	3.96	7	0.57	0.27	1.19	1.55	6	0.26	0.12	0.57	5.51	13	0.42	0.25	0.73
>90	1.42	4	0.36	0.13	0.95	4.13	9	0.46	0.24	0.88	5.56	13	0.43	0.25	0.74
(4) Using the upper 95% prediction interval bound prevalence, univariate sensitivity and univariate specificity															
>40	20.17	12	1.68	0.95	2.96	1.80	1	1.80	0.25	12.78	21.97	13	1.69	0.98	2.91
>50	19.48	12	1.62	0.92	2.86	1.82	1	1.82	0.26	12.90	21.30	13	1.64	0.95	2.82
>60	15.76	10	1.58	0.85	2.93	4.25	3	1.42	0.46	4.39	20.01	13	1.54	0.89	2.65
>65	13.99	9	1.55	0.81	2.99	5.52	4	1.38	0.52	3.68	19.52	13	1.50	0.87	2.59
>70	13.38	9	1.49	0.77	2.86	5.63	4	1.41	0.53	3.75	19.01	13	1.46	0.85	2.52
>72	12.93	9	1.44	0.75	2.76	5.67	4	1.42	0.53	3.78	18.60	13	1.43	0.83	2.46
>80	9.29	7	1.33	0.63	2.78	8.90	6	1.48	0.67	3.30	18.18	13	1.40	0.81	2.41
>90	3.17	4	0.79	0.30	2.11	19.51	9	2.17	1.13	4.17	22.68	13	1.74	1.01	3.00
(5) Using the studies observed prevalence, univariate sensitivity and univariate specificity															
>40	12.55	12	1.05	0.59	1.84	0.80	1	0.80	0.11	5.67	13.35	13	1.03	0.60	1.77
>50	12.65	12	1.05	0.60	1.86	0.80	1	0.80	0.11	5.67	13.45	13	1.03	0.60	1.78
>60	11.32	10	1.13	0.61	2.10	1.92	3	0.64	0.21	1.99	13.25	13	1.02	0.59	1.76
>65	10.26	9	1.14	0.59	2.19	2.54	4	0.64	0.24	1.69	12.80	13	0.98	0.57	1.70
>70	9.88	9	1.10	0.57	2.11	2.59	4	0.65	0.24	1.72	12.47	13	0.96	0.56	1.65
>72	9.72	9	1.08	0.56	2.07	2.60	4	0.65	0.24	1.73	12.32	13	0.95	0.55	1.63
>80	7.08	7	1.01	0.48	2.12	4.23	6	0.71	0.32	1.57	11.31	13	0.87	0.51	1.50
>90	2.47	4	0.62	0.23	1.64	10.48	9	1.16	0.61	2.24	12.95	13	1.00	0.58	1.72

Now, the Lo [142] study results are presented.

Table 5.7 shows the E/O ratio for the calibration of PPV and NPV calculated using the unstratified prevalence, sensitivity and specificity (approach 1); clearly these PPV and NPV dramatically over predicts the number of events, with an E/O of 2.16 for the 65% threshold. Using PPV and NPV based on Approach 2 (where the univariate prevalence, sensitivity and specificity are used) leads to a poorer calibration performance of the

predictive test with an E/O ratio of 2.38 at 65%. When using the lower bound of the 95% prediction interval for the prevalence (approach 3), the calibration performance of the predictive test (PPV and NPV) is much improved, with an E/O of 1.27 at the 65% threshold, although it still over predicts.

Table 5.7: E/O ratios using 5 approaches for Lo study at the 0-20 minute's time-point

Approach	No. of hypocalcaemia who tested positive			95% CI		No. of hypocalcaemia who tested negative			95% CI		Total with hypocalcaemia			95% CI	
Threshold	E	O	E/O	Lower	Upper	E	O	E/O	Lower	Upper	E	O	E/O	Lower	Upper
(1)	Using the Unstratified prevalence, Unstratified sensitivity and Unstratified specificity														
>40	21.70	11	1.97	1.09	3.56	1.12	0				22.82	11	2.07	1.15	3.75
>50	19.39	11	1.76	0.98	3.18	1.39	0				20.78	11	1.89	1.05	3.41
>60	20.63	11	1.88	1.04	3.39	2.83	0				23.46	11	2.13	1.18	3.85
>65	20.13	11	1.83	1.01	3.30	3.62	0				23.75	11	2.16	1.20	3.90
>70	19.83	11	1.80	1.00	3.26	3.69	0				23.52	11	2.14	1.18	3.86
>72	20.27	11	1.84	1.02	3.33	3.64	0				23.91	11	2.17	1.20	3.93
>80	15.12	10	1.51	0.81	2.81	6.64	1	6.64	0.93	47.11	21.76	11	1.98	1.10	3.57
>90	10.75	10	1.08	0.58	2.00	13.14	1	13.14	1.85	93.27	23.89	11	2.17	1.20	3.92
(2)	Using the univariate prevalence, univariate sensitivity and univariate specificity														
>40	22.80	11	2.07	1.15	3.74	1.08	0				23.88	11	2.17	1.20	3.92
>50	19.67	11	1.79	0.99	3.23	1.41	0				21.09	11	1.92	1.06	3.46
>60	22.99	11	2.09	1.16	3.77	2.72	0				25.71	11	2.34	1.29	4.22
>65	22.64	11	2.06	1.14	3.72	3.49	0				26.14	11	2.38	1.32	4.29
>70	21.99	11	2.00	1.11	3.61	3.59	0				25.58	11	2.33	1.29	4.20
>72	22.94	11	2.09	1.15	3.77	3.51	0				26.46	11	2.41	1.33	4.34
>80	15.26	10	1.53	0.82	2.84	6.40	1	6.40	0.90	45.45	21.66	11	1.97	1.09	3.56
>90	9.48	10	0.95	0.51	1.76	15.58	1	15.58	2.19	110.62	25.06	11	2.28	1.26	4.11
(3)	Using the lower 95% prediction interval bound prevalence, univariate sensitivity and univariate specificity														
>40	10.84	11	0.99	0.55	1.78	0.42	0				11.26	11	1.02	0.57	1.85
>50	9.70	11	0.88	0.49	1.59	0.55	0				10.24	11	0.93	0.52	1.68
>60	12.52	11	1.14	0.63	2.06	1.07	0				13.59	11	1.24	0.68	2.23
>65	12.62	11	1.15	0.64	2.07	1.38	0				14.00	11	1.27	0.70	2.30
>70	12.35	11	1.12	0.62	2.03	1.42	0				13.77	11	1.25	0.69	2.26
>72	13.16	11	1.20	0.66	2.16	1.39	0				14.55	11	1.32	0.73	2.39
>80	8.91	10	0.89	0.48	1.66	2.57	1	2.57	0.36	18.24	11.48	11	1.04	0.58	1.88
>90	5.69	10	0.57	0.31	1.06	6.68	1	6.68	0.94	47.41	12.36	11	1.12	0.62	2.03
(4)	Using the upper 95% prediction interval bound prevalence, univariate sensitivity and univariate specificity														
>40	39.21	11	3.56	1.97	6.44	2.70	0				41.91	11	3.81	2.11	6.88
>50	32.27	11	2.93	1.62	5.30	3.56	0				35.83	11	3.26	1.80	5.88
>60	33.67	11	3.06	1.69	5.53	6.62	0				40.29	11	3.66	2.03	6.61
>65	32.41	11	2.95	1.63	5.32	8.36	0				40.77	11	3.71	2.05	6.69
>70	31.22	11	2.84	1.57	5.13	8.59	0				39.81	11	3.62	2.00	6.54
>72	31.94	11	2.90	1.61	5.24	8.43	0				40.37	11	3.67	2.03	6.63
>80	20.90	10	2.09	1.12	3.88	14.76	1	14.76	2.08	104.79	35.66	11	3.24	1.80	5.85
>90	12.69	10	1.27	0.68	2.36	31.52	1	31.52	4.44	223.75	44.21	11	4.02	2.23	7.26
(5)	Using the studies observed prevalence, univariate sensitivity and univariate specificity														
>40	12.50	11	1.14	0.63	2.05	0.50	0				12.99	11	1.18	0.65	2.13
>50	11.12	11	1.01	0.56	1.83	0.65	0				11.77	11	1.07	0.59	1.93
>60	14.15	11	1.29	0.71	2.32	1.26	0				15.41	11	1.40	0.78	2.53
>65	14.21	11	1.29	0.72	2.33	1.63	0				15.84	11	1.44	0.80	2.60
>70	13.89	11	1.26	0.70	2.28	1.67	0				15.56	11	1.41	0.78	2.55
>72	14.75	11	1.34	0.74	2.42	1.64	0				16.39	11	1.49	0.83	2.69
>80	9.96	10	1.00	0.54	1.85	3.02	1	3.02	0.43	21.46	12.98	11	1.18	0.65	2.13
>90	6.33	10	0.63	0.34	1.18	7.80	1	7.80	1.10	55.40	14.13	11	1.28	0.71	2.32

Where blank, this is due to observed cases being 0 so an E/O ratio was not possible

In approach 5, when using the studies observed prevalence of 0.11 and combining that with the univariate meta-analysis values for sensitivity and specificity to derive PPV

and NPV, the calibration performance of the predictive test is closer to 1 than when using PPV and NPV derived using approaches 1 and 2, as the 65% threshold yields an E/O of 1.44.

This analysis shows again that using the unstratified or average univariate meta-analysis results for sensitivity, specificity and prevalence to derive PPV and NPV in this particular study does not give accurate predictions for clinical use as the calibration is poor. A key reason for this is that the average prevalence across all studies (0.21) is very different to this settings prevalence (0.11). When using the study's observed prevalence, or indeed the similar lower bound of the predicted interval for prevalence, the calibration is markedly improved. However, there is still some over-prediction in the performance of the predictive test (PPV and NPV), which may be due to chance (small number of events and patients) but may also be due to additional heterogeneity in the sensitivity and specificity for this study compared to the averages used. This is evaluated in a formal meta-analysis of the E/O statistics across studies later in section 5.4.3.

Now the Lombardi [143] study results are presented.

Table 5.8 shows the E/O ratio calculated using the unstratified prevalence, sensitivity and specificity (approach 1). When this approach is used to derive PPV and NPV, it is found the number of events is under-predicted with an E/O of 0.62 for the 65% threshold. The use of the second approach (where the univariate prevalence, sensitivity and specificity are used) to derive PPV and NPV, results in slightly better calibration

performance, with an E/O ratio of 0.68 at 65%. When using the upper bound of the 95% prediction interval for the prevalence (approach 4), an even better calibration performance (of PPV and NPV) is obtained compared to the use of the first two approaches to derive PPV and NPV, with an E/O of 1.10 at the 65% threshold.

In approach 5, when PPV and NPV are derived using the study's observed prevalence of 0.31 and combining that with the univariate meta-analysis values for sensitivity and specificity, the calibration performance (E/O ratio) is closer to 1 than compared to the calibration performance of PPV and NPV derived from approaches 1 and 2, as the 65% threshold yields an E/O of 0.88. This again indicates that when using the study's own prevalence, calibration performance is more accurate for this study, as the study's observed prevalence of 0.31 is much higher than the average study prevalence of 0.21.

Table 5.8: E/O ratios using 5 approaches for Lombardi study at the 0-20 minute's time-point

Approach	No. of hypocalcaemia who tested positive			95% CI		No. of hypocalcaemia who tested negative			95% CI		Total with hypocalcaemia			95% CI	
Threshold	E	O	E/O	Lower	Upper	E	O	E/O	Lower	Upper	E	O	E/O	Lower	Upper
(1)	Using the Unstratified prevalence, Unstratified sensitivity and Unstratified specificity														
>40	8.37	15	0.56	0.34	0.93	0.90	1	0.90	0.13	6.37	9.27	16	0.58	0.35	0.95
>50	9.15	15	0.61	0.37	1.01	0.77	1	0.77	0.11	5.45	9.91	16	0.62	0.38	1.01
>60	8.78	14	0.63	0.37	1.06	1.66	2	0.83	0.21	3.31	10.43	16	0.65	0.40	1.06
>65	7.78	13	0.60	0.35	1.03	2.20	3	0.73	0.24	2.27	9.98	16	0.62	0.38	1.02
>70	8.03	13	0.62	0.36	1.06	2.16	3	0.72	0.23	2.23	10.19	16	0.64	0.39	1.04
>72	7.72	13	0.59	0.34	1.02	2.20	3	0.73	0.24	2.27	9.92	16	0.62	0.38	1.01
>80	7.28	11	0.66	0.37	1.20	3.45	5	0.69	0.29	1.66	10.73	16	0.67	0.41	1.10
>90	1.34	2	0.67	0.17	2.69	7.66	14	0.55	0.32	0.92	9.01	16	0.56	0.34	0.92
(2)	Using the univariate prevalence, univariate sensitivity and univariate specificity														
>40	8.79	15	0.59	0.35	0.97	0.87	1	0.87	0.12	6.15	9.66	16	0.60	0.37	0.99
>50	9.28	15	0.62	0.37	1.03	0.78	1	0.78	0.11	5.56	10.06	16	0.63	0.39	1.03
>60	9.78	14	0.70	0.41	1.18	1.59	2	0.80	0.20	3.18	11.37	16	0.71	0.44	1.16
>65	8.75	13	0.67	0.39	1.16	2.12	3	0.71	0.23	2.19	10.87	16	0.68	0.42	1.11
>70	8.90	13	0.68	0.40	1.18	2.10	3	0.70	0.23	2.18	11.00	16	0.69	0.42	1.12
>72	8.74	13	0.67	0.39	1.16	2.12	3	0.71	0.23	2.19	10.86	16	0.68	0.42	1.11
>80	7.35	11	0.67	0.37	1.21	3.33	5	0.67	0.28	1.60	10.68	16	0.67	0.41	1.09
>90	1.19	2	0.59	0.15	2.37	9.09	14	0.65	0.38	1.10	10.27	16	0.64	0.39	1.05
(3)	Using the lower 95% prediction interval bound prevalence, univariate sensitivity and univariate specificity														
>40	4.18	15	0.28	0.17	0.46	0.34	1	0.34	0.05	2.39	4.52	16	0.28	0.17	0.46
>50	4.58	15	0.31	0.18	0.51	0.30	1	0.30	0.04	2.14	4.88	16	0.30	0.19	0.50
>60	5.33	14	0.38	0.23	0.64	0.62	2	0.31	0.08	1.25	5.95	16	0.37	0.23	0.61
>65	4.88	13	0.38	0.22	0.65	0.84	3	0.28	0.09	0.86	5.71	16	0.36	0.22	0.58
>70	5.00	13	0.38	0.22	0.66	0.83	3	0.28	0.09	0.86	5.83	16	0.36	0.22	0.59
>72	5.01	13	0.39	0.22	0.66	0.84	3	0.28	0.09	0.86	5.85	16	0.37	0.22	0.60
>80	4.29	11	0.39	0.22	0.70	1.34	5	0.27	0.11	0.64	5.63	16	0.35	0.22	0.57
>90	0.71	2	0.36	0.09	1.42	3.90	14	0.28	0.16	0.47	4.61	16	0.29	0.18	0.47
(4)	Using the upper 95% prediction interval bound prevalence, univariate sensitivity and univariate specificity														
>40	15.13	15	1.01	0.61	1.67	2.16	1	2.16	0.30	15.33	17.29	16	1.08	0.66	1.76

Approach	No. of hypocalcaemia who tested positive			95% CI		No. of hypocalcaemia who tested negative			95% CI		Total with hypocalcaemia			95% CI	
Threshold	E	O	E/O	Lower	Upper	E	O	E/O	Lower	Upper	E	O	E/O	Lower	Upper
>50	15.22	15	1.01	0.61	1.68	1.97	1	1.97	0.28	13.97	17.19	16	1.07	0.66	1.75
>60	14.33	14	1.02	0.61	1.73	3.87	2	1.94	0.48	7.74	18.20	16	1.14	0.70	1.86
>65	12.52	13	0.96	0.56	1.66	5.08	3	1.69	0.55	5.25	17.60	16	1.10	0.67	1.80
>70	12.64	13	0.97	0.56	1.67	5.04	3	1.68	0.54	5.20	17.67	16	1.10	0.68	1.80
>72	12.17	13	0.94	0.54	1.61	5.09	3	1.70	0.55	5.26	17.26	16	1.08	0.66	1.76
>80	10.06	11	0.91	0.51	1.65	7.68	5	1.54	0.64	3.69	17.75	16	1.11	0.68	1.81
>90	1.59	2	0.79	0.20	3.17	18.38	14	1.31	0.78	2.22	19.97	16	1.25	0.76	2.04
(5)	Using the studies observed prevalence, univariate sensitivity and univariate specificity														
>40	11.84	15	0.79	0.48	1.31	1.37	1	1.37	0.19	9.75	13.22	16	0.83	0.51	1.35
>50	12.22	15	0.81	0.49	1.35	1.24	1	1.24	0.18	8.82	13.46	16	0.84	0.52	1.37
>60	12.15	14	0.87	0.51	1.47	2.50	2	1.25	0.31	4.99	14.65	16	0.92	0.56	1.49
>65	10.74	13	0.83	0.48	1.42	3.30	3	1.10	0.36	3.42	14.04	16	0.88	0.54	1.43
>70	10.88	13	0.84	0.49	1.44	3.28	3	1.09	0.35	3.39	14.16	16	0.88	0.54	1.44
>72	10.57	13	0.81	0.47	1.40	3.31	3	1.10	0.36	3.42	13.88	16	0.87	0.53	1.42
>80	8.81	11	0.80	0.44	1.45	5.11	5	1.02	0.43	2.46	13.92	16	0.87	0.53	1.42
>90	1.40	2	0.70	0.18	2.81	13.20	14	0.94	0.56	1.59	14.60	16	0.91	0.56	1.49

The results of calibration for 0-20 minutes for other studies are shown in APPENDIX B and reveal similar findings.

5.3.2.2 1-2 hours

Now consider the time-point of 1-2 hours and the E/O ratio when the patients are pooled together across all studies in Table 5.9 and treated as one big dataset.

Table 5.9: E/O ratios for all the patients combined using the unstratified and univariate meta-analysis approaches at the 1-2 hours' time-point

% PTH Decrease	No. of hypocalcaemia who tested positive			95% CI		No. of hypocalcaemia who tested negative			95% CI		Total with hypocalcaemia			95% CI	
	E	O	E/O	Lower	Upper	E	O	E/O	Lower	Upper	E	O	E/O	Lower	Upper
<i>Approach 1</i>															
>40	34.54	44	0.79	0.58	1.05	1.55	3	0.52	0.17	1.60	36.09	47	0.77	0.58	1.02
>50	35.35	44	0.80	0.60	1.08	1.55	3	0.52	0.17	1.60	36.91	47	0.79	0.59	1.05
>60	37.36	40	0.93	0.69	1.27	1.55	7	0.22	0.11	0.46	38.91	47	0.83	0.62	1.10
>65	36.56	38	0.96	0.70	1.32	2.33	9	0.26	0.13	0.50	38.90	47	0.83	0.62	1.10
>70	37.90	38	1.00	0.73	1.37	2.33	9	0.26	0.13	0.50	40.23	47	0.86	0.64	1.14
>72	36.42	38	0.96	0.70	1.32	3.74	9	0.42	0.22	0.80	40.16	47	0.85	0.64	1.14
>80	31.98	32	1.00	0.71	1.41	7.91	15	0.53	0.32	0.87	39.89	47	0.85	0.64	1.13
>90	25.00	16	1.56	0.96	2.55	15.96	31	0.51	0.36	0.73	40.96	47	0.87	0.65	1.16
<i>Approach 2</i>															
>40	35.07	44	0.80	0.59	1.07	1.40	3	0.47	0.15	1.45	36.47	47	0.78	0.58	1.03
>50	35.87	44	0.82	0.61	1.10	1.41	3	0.47	0.15	1.45	37.28	47	0.79	0.60	1.06
>60	37.74	40	0.94	0.69	1.29	1.40	7	0.20	0.10	0.42	39.14	47	0.83	0.63	1.11
>65	37.77	38	0.99	0.72	1.37	2.09	9	0.23	0.12	0.45	39.85	47	0.85	0.64	1.13
>70	38.18	38	1.00	0.73	1.38	2.11	9	0.23	0.12	0.45	40.28	47	0.86	0.64	1.14
>72	36.62	38	0.96	0.70	1.32	3.54	9	0.39	0.20	0.76	40.16	47	0.85	0.64	1.14
>80	32.81	32	1.03	0.73	1.45	7.16	15	0.48	0.29	0.79	39.98	47	0.85	0.64	1.13
>90	25.00	16	1.56	0.96	2.55	14.60	31	0.47	0.33	0.67	39.60	47	0.84	0.63	1.12

When using the unstratified approach to derive PPV and NPV, Table 5.9 shows that the E/O ratio combined for predicting those who get hypocalcaemia is less than 1

consistently. When using the univariate approach to derive PPV and NPV, similar results are found with the E/O ratio being under 1.

Now, each study on its own will be assessed. As in the 0-20 minute's analysis, the calibration performance of the predictive test (PPV and NPV) in each study is improved when using the study's observed prevalence (approach (5)).

Table 5.10: E/O ratios using 5 approaches for McLeod study at the 1-2 hours' time-point

Approach	No. of hypocalcaemia who tested positive			95% CI		No. of hypocalcaemia who tested negative			95% CI		Total with hypocalcaemia			95% CI	
Threshold	E	O	E/O	Lower	Upper	E	O	E/O	Lower	Upper	E	O	E/O	Lower	Upper
(1) Using the Unstratified prevalence, Unstratified sensitivity and Unstratified specificity															
>40	9.77	12	0.81	0.46	1.43	0.20	0				9.97	12	0.83	0.47	1.46
>50	9.46	12	0.79	0.45	1.39	0.26	0				9.71	12	0.81	0.46	1.43
>60	10.76	12	0.90	0.51	1.58	0.26	0				11.02	12	0.92	0.52	1.62
>65	9.71	11	0.88	0.49	1.59	0.43	1	0.43	0.06	3.07	10.14	12	0.85	0.48	1.49
>70	10.83	11	0.98	0.55	1.78	0.42	1	0.42	0.06	2.97	11.25	12	0.94	0.53	1.65
>72	10.51	11	0.96	0.53	1.72	0.68	1	0.68	0.10	4.81	11.18	12	0.93	0.53	1.64
>80	9.14	9	1.02	0.53	1.95	1.50	3	0.50	0.16	1.55	10.64	12	0.89	0.50	1.56
>90	7.00	7	1.00	0.48	2.10	3.20	5	0.64	0.27	1.54	10.20	12	0.85	0.48	1.50
(2) Using the univariate prevalence, univariate sensitivity and univariate specificity															
>40	9.92	12	0.83	0.47	1.46	0.21	0				10.12	12	0.84	0.48	1.49
>50	9.59	12	0.80	0.45	1.41	0.26	0				9.86	12	0.82	0.47	1.45
>60	10.86	12	0.91	0.51	1.59	0.27	0				11.13	12	0.93	0.53	1.63
>65	10.03	11	0.91	0.51	1.65	0.44	1	0.44	0.06	3.10	10.47	12	0.87	0.50	1.54
>70	10.91	11	0.99	0.55	1.79	0.43	1	0.43	0.06	3.04	11.34	12	0.94	0.54	1.66
>72	10.56	11	0.96	0.53	1.73	0.72	1	0.72	0.10	5.14	11.29	12	0.94	0.53	1.66
>80	9.38	9	1.04	0.54	2.00	1.53	3	0.51	0.16	1.58	10.90	12	0.91	0.52	1.60
>90	7.00	7	1.00	0.48	2.10	3.26	5	0.65	0.27	1.57	10.26	12	0.86	0.49	1.51
(3) Using the lower 95% prediction interval bound prevalence, univariate sensitivity and univariate specificity															
>40	4.82	12	0.40	0.23	0.71	0.08	0				4.90	12	0.41	0.23	0.72
>50	4.91	12	0.41	0.23	0.72	0.10	0				5.01	12	0.42	0.24	0.74
>60	6.39	12	0.53	0.30	0.94	0.10	0				6.49	12	0.54	0.31	0.95
>65	6.00	11	0.55	0.30	0.99	0.17	1	0.17	0.02	1.19	6.17	12	0.51	0.29	0.91
>70	7.17	11	0.65	0.36	1.18	0.16	1	0.16	0.02	1.16	7.34	12	0.61	0.35	1.08
>72	7.12	11	0.65	0.36	1.17	0.28	1	0.28	0.04	1.98	7.40	12	0.62	0.35	1.09
>80	6.90	9	0.77	0.40	1.47	0.60	3	0.20	0.06	0.62	7.50	12	0.63	0.36	1.10
>90	7.00	7	1.00	0.48	2.10	1.33	5	0.27	0.11	0.64	8.33	12	0.69	0.39	1.22
(4) Using the upper 95% prediction interval bound prevalence, univariate sensitivity and univariate specificity															
>40	16.55	12	1.38	0.78	2.43	0.52	0				17.08	12	1.42	0.81	2.51
>50	15.03	12	1.25	0.71	2.21	0.68	0				15.71	12	1.31	0.74	2.31
>60	14.80	12	1.23	0.70	2.17	0.69	0				15.49	12	1.29	0.73	2.27
>65	13.46	11	1.22	0.68	2.21	1.11	1	1.11	0.16	7.88	14.57	12	1.21	0.69	2.14
>70	13.59	11	1.24	0.68	2.23	1.09	1	1.09	0.15	7.75	14.69	12	1.22	0.70	2.15
>72	12.94	11	1.18	0.65	2.12	1.81	1	1.81	0.26	12.85	14.75	12	1.23	0.70	2.16
>80	10.85	9	1.21	0.63	2.32	3.66	3	1.22	0.39	3.78	14.50	12	1.21	0.69	2.13
>90	7.00	7	1.00	0.48	2.10	7.31	5	1.46	0.61	3.51	14.31	12	1.19	0.68	2.10
(5) Using the studies observed prevalence, univariate sensitivity and univariate specificity															
>40	10.58	12	0.88	0.50	1.55	0.23	0				10.81	12	0.90	0.51	1.59
>50	10.17	12	0.85	0.48	1.49	0.29	0				10.46	12	0.87	0.50	1.54
>60	11.34	12	0.94	0.54	1.66	0.30	0				11.63	12	0.97	0.55	1.71
>65	10.45	11	0.95	0.53	1.72	0.48	1	0.48	0.07	3.42	10.93	12	0.91	0.52	1.60
>70	11.26	11	1.02	0.57	1.85	0.47	1	0.47	0.07	3.35	11.73	12	0.98	0.56	1.72
>72	10.88	11	0.99	0.55	1.79	0.80	1	0.80	0.11	5.67	11.67	12	0.97	0.55	1.71
>80	9.58	9	1.06	0.55	2.05	1.68	3	0.56	0.18	1.73	11.26	12	0.94	0.53	1.65
>90	7.00	7	1.00	0.48	2.10	3.57	5	0.71	0.30	1.72	10.57	12	0.88	0.50	1.55

Where blank, this is due to observed cases being 0 so an E/O ratio was not possible

For example, Table 5.10 shows the calibration performance (E/O ratio) for the McLeod [144] study and at a 65% threshold, when using the observed prevalence along with univariate sensitivity and univariate specificity to derive PPV and NPV (approach (5)) the E/O ratio is 0.91, but 0.87 using the univariate meta-analysis values to derive PPV and NPV (approach (2)) and 0.85 when using the unstratified values for prevalence, sensitivity and specificity to derive PPV and NPV (approach (1)).

Table 5.11: E/O ratios using 5 approaches for Lam study at the 1-2 hours' time-point

Approach	No. of hypocalcaemia who tested positive			95% CI		No. of hypocalcaemia who tested negative			95% CI		Total with hypocalcaemia			95% CI	
Threshold	E	O	E/O	Lower	Upper	E	O	E/O	Lower	Upper	E	O	E/O	Lower	Upper
(1)	Using the Unstratified prevalence, Unstratified sensitivity and Unstratified specificity														
>40	7.33	12	0.61	0.35	1.08	0.38	0				7.71	12	0.64	0.36	1.13
>50	6.99	12	0.58	0.33	1.03	0.39	0				7.38	12	0.61	0.35	1.08
>60	7.93	12	0.66	0.38	1.16	0.35	0				8.28	12	0.69	0.39	1.21
>65	8.00	12	0.67	0.38	1.17	0.52	0				8.52	12	0.71	0.40	1.25
>70	9.47	12	0.79	0.45	1.39	0.48	0				9.95	12	0.83	0.47	1.46
>72	9.81	12	0.82	0.46	1.44	0.74	0				10.54	12	0.88	0.50	1.55
>80	8.38	11	0.76	0.42	1.38	1.62	1	1.62	0.23	11.50	10.00	12	0.83	0.47	1.47
>90	8.00	8	1.00	0.50	2.00	3.20	4	0.80	0.30	2.13	11.20	12	0.93	0.53	1.64
(2)	Using the univariate prevalence, univariate sensitivity and univariate specificity														
>40	7.44	12	0.62	0.35	1.09	0.39	0				7.83	12	0.65	0.37	1.15
>50	7.09	12	0.59	0.34	1.04	0.40	0				7.49	12	0.62	0.35	1.10
>60	8.01	12	0.67	0.38	1.17	0.36	0				8.37	12	0.70	0.40	1.23
>65	8.26	12	0.69	0.39	1.21	0.52	0				8.78	12	0.73	0.42	1.29
>70	9.54	12	0.80	0.45	1.40	0.49	0				10.03	12	0.84	0.47	1.47
>72	9.86	12	0.82	0.47	1.45	0.79	0				10.65	12	0.89	0.50	1.56
>80	8.59	11	0.78	0.43	1.41	1.65	1	1.65	0.23	11.69	10.24	12	0.85	0.48	1.50
>90	8.00	8	1.00	0.50	2.00	3.26	4	0.82	0.31	2.17	11.26	12	0.94	0.53	1.65
(3)	Using the lower 95% prediction interval bound prevalence, univariate sensitivity and univariate specificity														
>40	3.62	12	0.30	0.17	0.53	0.15	0				3.77	12	0.31	0.18	0.55
>50	3.63	12	0.30	0.17	0.53	0.15	0				3.78	12	0.32	0.18	0.55
>60	4.71	12	0.39	0.22	0.69	0.14	0				4.84	12	0.40	0.23	0.71
>65	4.94	12	0.41	0.23	0.73	0.20	0				5.15	12	0.43	0.24	0.76
>70	6.27	12	0.52	0.30	0.92	0.19	0				6.46	12	0.54	0.31	0.95
>72	6.64	12	0.55	0.31	0.97	0.30	0				6.95	12	0.58	0.33	1.02
>80	6.33	11	0.58	0.32	1.04	0.65	1	0.65	0.09	4.60	6.98	12	0.58	0.33	1.02
>90	8.00	8	1.00	0.50	2.00	1.33	4	0.33	0.12	0.89	9.33	12	0.78	0.44	1.37
(4)	Using the upper 95% prediction interval bound prevalence, univariate sensitivity and univariate specificity														
>40	12.42	12	1.03	0.59	1.82	0.99	0				13.40	12	1.12	0.63	1.97
>50	11.11	12	0.93	0.53	1.63	1.01	0				12.13	12	1.01	0.57	1.78
>60	10.90	12	0.91	0.52	1.60	0.92	0				11.83	12	0.99	0.56	1.74
>65	11.08	12	0.92	0.52	1.63	1.33	0				12.41	12	1.03	0.59	1.82
>70	11.89	12	0.99	0.56	1.75	1.25	0				13.14	12	1.10	0.62	1.93
>72	12.08	12	1.01	0.57	1.77	1.98	0				14.05	12	1.17	0.66	2.06
>80	9.94	11	0.90	0.50	1.63	3.95	1	3.95	0.56	28.02	13.89	12	1.16	0.66	2.04
>90	8.00	8	1.00	0.50	2.00	7.31	4	1.83	0.69	4.87	15.31	12	1.28	0.72	2.25
(5)	Using the studies observed prevalence, univariate sensitivity and univariate specificity														
>40	9.87	12	0.82	0.47	1.45	0.62	0				10.49	12	0.87	0.50	1.54
>50	9.12	12	0.76	0.43	1.34	0.63	0				9.75	12	0.81	0.46	1.43
>60	9.57	12	0.80	0.45	1.40	0.58	0				10.15	12	0.85	0.48	1.49
>65	9.79	12	0.82	0.46	1.44	0.84	0				10.63	12	0.89	0.50	1.56
>70	10.87	12	0.91	0.51	1.59	0.78	0				11.65	12	0.97	0.55	1.71
>72	11.11	12	0.93	0.53	1.63	1.25	0				12.37	12	1.03	0.59	1.81
>80	9.38	11	0.85	0.47	1.54	2.57	1	2.57	0.36	18.23	11.95	12	1.00	0.57	1.75
>90	8.00	8	1.00	0.50	2.00	4.95	4	1.24	0.46	3.30	12.95	12	1.08	0.61	1.90

Where blank, this is due to observed cases being 0 so an E/O ratio was not possible

Similarly, in the Lam [141] study E/O is 0.71, 0.73 and 0.89 when using the unstratified (approach (1)), univariate meta-analysis average (approach (2)), or observed study prevalence (approach (5)) to derive PPV and NPV (Table 5.11) to assess calibration performance of the predictive test. Results for other studies are shown in APPENDIX B.

Similar conclusions were found at 6 hours (see APPENDIX B).

5.3.3 Meta-analysis of E/O estimates

5.3.3.1 0-20 minutes time-point

To display the range of E/O values across studies at each threshold and to examine if the variation is due to chance or heterogeneity, a meta-analysis is possible. Now take the threshold of 65% as an example, and consider a meta-analysis of the calibration performance of PPV and NPV derived from approaches 1, 2 and 5. Approaches 3 and 4 are not considered here as they are extreme examples due to the low and high values of prevalence used in those approaches. Approaches 1, 2 and 5 represent a more realistic approach in practice.

(i) Approach 1:

This approach used the unstratified prevalence, sensitivity and specificity to derive PPV and NPV for the predictive test.

Figure 5.1 shows that although the average E/O value is at 1.02, 95% CI: (0.56, 1.86), the values for individual studies differs as some are lower than 1 and some are above 1. This is reflected in the heterogeneity as I-squared=85.5%, which is very high, with

statistical significance of heterogeneity in E/O ($p < 0.001$). The 95% prediction interval as shown in Figure 5.1 is very wide reflecting the uncertainty in the E/O value when the unstratified approach is used to derive PPV and NPV to any one setting.

(ii) Approach 2:

This approach used the univariate prevalence, univariate sensitivity and univariate specificity to derive PPV and NPV for the predictive test.

Figure 5.1 shows that, although the average E/O value is at 1.11 (0.61, 2.03), the values for individual studies still differ as some are lower than 1 and some are above 1. This is reflected in the heterogeneity as $I^2 = 85.7\%$ which is again very high, with statistical significance ($p < 0.001$). The 95% prediction interval as shown in Figure 5.1 is also very wide reflecting in the uncertainty surrounding E/O in a single setting. Thus, using PPV and NPV as derived from either approaches 1 and 2 appears to calibrate well on average across all studies, but the wide prediction interval and large heterogeneity indicate potentially poorer calibration performance in individual clinical settings.

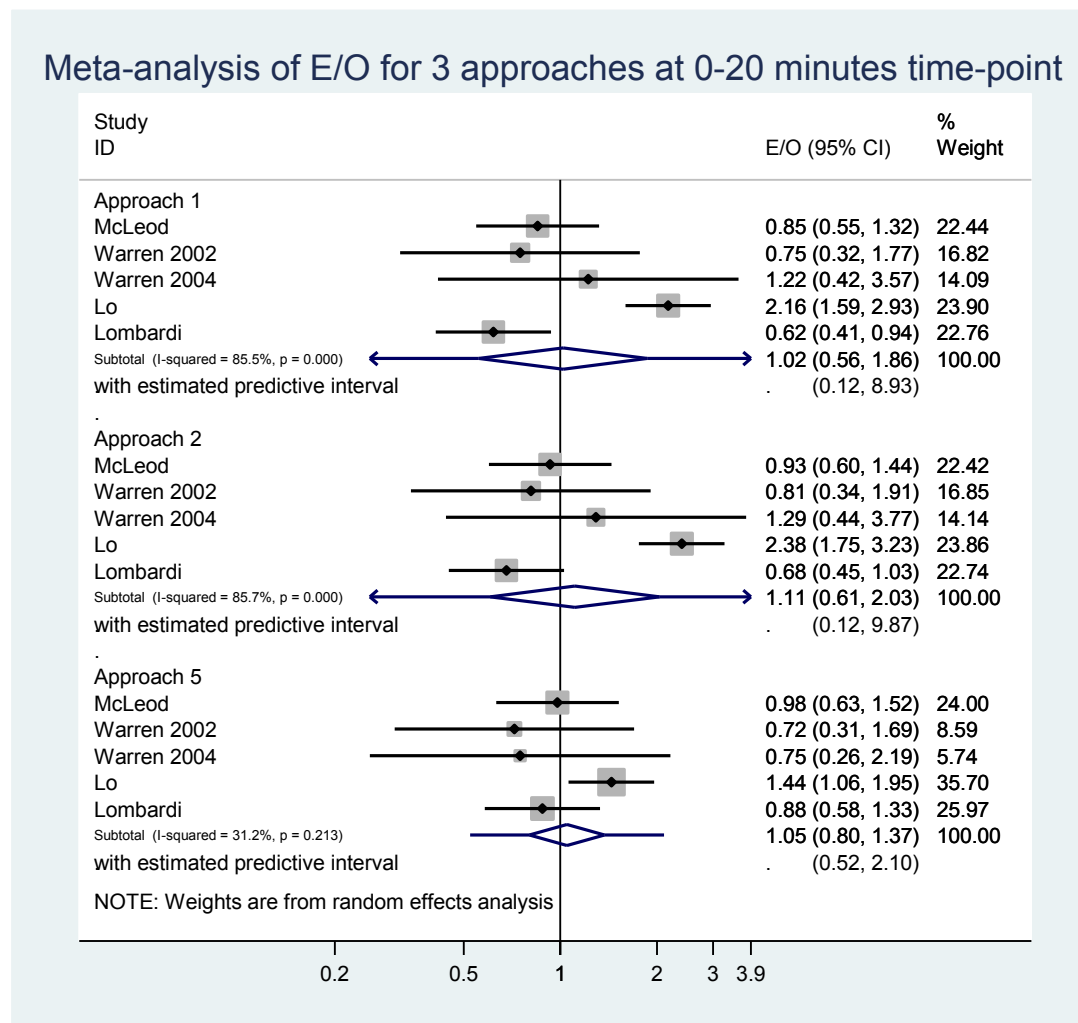
(iii) Approach 5:

This approach uses the study's observed prevalence, univariate sensitivity and univariate specificity to derive the PPV and NPV of the predictive test.

Figure 5.1 shows that the average E/O value is at 1.05 (0.80, 1.37), and now the E/O values for individual studies are quite similar to this. This is reflected in very little heterogeneity as $I^2 = 31.2\%$ and $p\text{-value} = 0.21$. Thus by using the observed

prevalence in each study, the heterogeneity in E/O statistics is reduced drastically and any variation across studies appears potentially due to chance. Further the summary E/O of 1.05 indicates the calibration performance of the predictive test is very good, though the 95% CI is wide. Of particular interest are the Lo [142] and Lombardi [143] studies; using the first two approaches to derive PPV and NPV, the calibration performance is shown to be over and under predicting greatly, whilst using fifth approach to derive PPV and NPV has much better calibration performance with the E/O much closer to 1 (Figure 5.1).

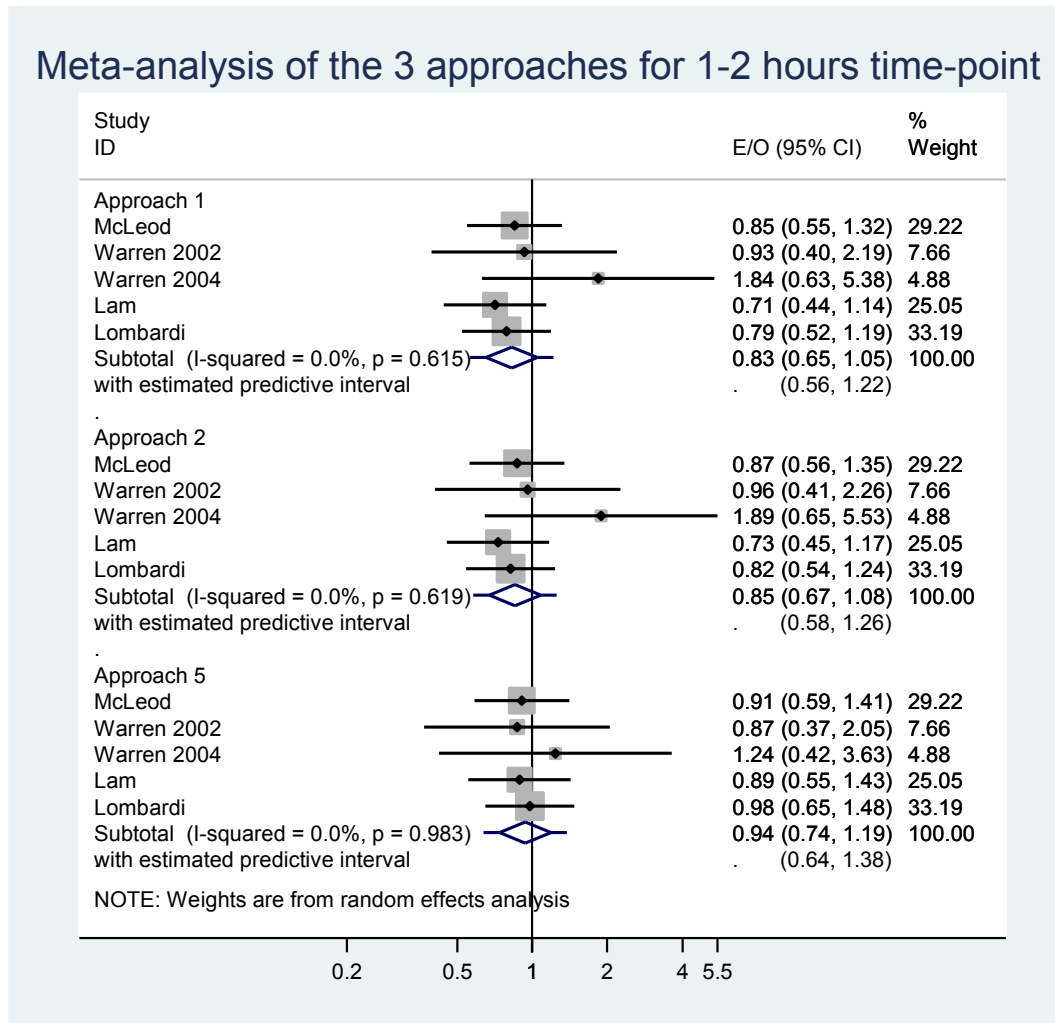
Figure 5.1: Meta-analysis of E/O for the 3 different approaches at the 0-20 minute's time-point



5.3.3.2 1-2 hours' time-point

The same finding can be seen in the meta-analysis at 1-2 hours (Figure 5.2). Although there is no heterogeneity in any analysis (I-squared=0%), the approach utilising the study's own prevalence to derive PPV and NPV improves the calibration performance of the predictive test with an E/O of 0.94, closer to 1 than the unstratified (0.83) and univariate (0.85) approaches.

Figure 5.2: Meta-analysis of E/O for the 3 different approaches at the 1-2 hours' time-point



5.3.3.3 Meta-analysis results for Approaches 1, 2 and 5 for the thresholds 60-80

In sections 5.4.3.1 and 5.4.3.2, threshold 65% was used for meta-analysis, in terms of the E/O ratio across the studies at the two time-points of 0-20 minutes and 1-2 hours (as 6 hours only had 2 studies contributing data, this was not meta-analysed). Now presented are the meta-analysis results from other thresholds 60-80% to show a general overview in Table 5.12.

For the 0-20 minute's time-point, in general deriving PPV and NPV using approach 1 gives a good average effect around 1.00 of the calibration performance of the predictive test, but also has high heterogeneity with I-squared being around 60%. Deriving PPV and NPV using approach 2 gives the same level of heterogeneity but with a slightly higher average effect around 1.09 of the calibration performance. Deriving PPV and NPV using approach 5 has a good average effect around 0.99 of the calibration performance, apart from the 80% threshold which has an E/O of 0.92, and there is no heterogeneity present which shows deriving PPV and NPV using the studies observed prevalence combined with the univariate values for sensitivity and specificity eliminates between-study heterogeneity.

For the 1-2 hours' time-point, the results are interesting, as there is no between-study heterogeneity present in the calibration performance of PPV and NPV derived from any of the approaches used. The calibration performance of PPV and NPV derived from approach 1 has an average E/O around 0.85, with the calibration performance of PPV and NPV derived from approach 2 having similar E/O ratios but slightly higher. In this example, the calibration performance (E/O) improves for PPV and NPV derived from approach 5 compared to the calibration performance of PPV and NPV derived from the first two approaches, with the E/O being around 0.95. Although there is no between-study heterogeneity in the calibration performance of the PPV and NPV derived from the first two approaches, the calibration is poorer compared to the PPV and NPV derived from approach 5 as it is almost 0.1 lower, which could make an important difference in clinical practice.

Table 5.12: Meta-analysis values for E/O for various thresholds at the 0-20 minutes and 1-2 hour time-points

	Pooled	95% CI		I ²	τ^2	95% PI	
0-20minutes	E/O	Lower	Upper			Lower	Upper
<i>Approach (1)</i>							
>60	0.99	0.61	1.61	60.0%	0.17	0.21	4.61
>65	1.00	0.60	1.66	63.3%	0.20	0.19	5.13
>70	1.00	0.62	1.64	61.0%	0.18	0.21	4.81
>72	1.00	0.60	1.66	63.7%	0.20	0.19	5.19
>80	1.03	0.65	1.63	55.4%	0.14	0.25	4.23
<i>Approach (2)</i>							
>60	1.08	0.66	1.75	60.5%	0.17	0.23	5.06
>65	1.09	0.65	1.81	63.9%	0.20	0.21	5.65
>70	1.08	0.66	1.78	61.8%	0.19	0.22	5.31
>72	1.09	0.65	1.83	64.6%	0.21	0.20	5.84
>80	1.03	0.65	1.62	54.8%	0.14	0.25	4.14
<i>Approach (5)</i>							
>60	0.99	0.75	1.32	0%	0.00	0.63	1.58
>65	0.99	0.74	1.32	0%	0.00	0.62	1.57
>70	0.98	0.74	1.31	0%	0.00	0.62	1.56
>72	0.99	0.75	1.32	0%	0.00	0.62	1.58
>80	0.92	0.69	1.23	0%	0.00	0.58	1.47
<i>1-2 hours</i>							
<i>Approach (1)</i>							
>60	0.83	0.62	1.12	0%	0.00	0.52	1.34
>65	0.83	0.62	1.12	0%	0.00	0.52	1.35
>70	0.87	0.65	1.16	0%	0.00	0.54	1.40
>72	0.86	0.65	1.16	0%	0.00	0.54	1.39
>80	0.86	0.64	1.16	0%	0.00	0.53	1.39
<i>Approach (2)</i>							
>60	0.84	0.63	1.13	0%	0.00	0.53	1.35
>65	0.86	0.64	1.15	0%	0.00	0.54	1.38
>70	0.88	0.65	1.17	0%	0.00	0.55	1.41
>72	0.87	0.65	1.17	0%	0.00	0.54	1.40
>80	0.88	0.66	1.17	0%	0.00	0.55	1.41
<i>Approach (5)</i>							
>60	0.93	0.70	1.25	0%	0.00	0.58	1.50
>65	0.95	0.71	1.27	0%	0.00	0.59	1.52
>70	0.95	0.71	1.27	0%	0.00	0.59	1.52
>72	0.94	0.70	1.26	0%	0.00	0.59	1.51
>80	0.95	0.71	1.28	0%	0.00	0.59	1.53

5.3.4 Approach 6, 'internal-external' cross-validation

In Chapter 3, it was found that few articles performed external validation, with most preferring internal validation, but a few articles used the IECV approach. This approach will now be applied to the PTH data in relation to the calibration performance of PPV and NPV derived for using the predictive test in practice.

Here the IECV approach is used to examine if using the summary sensitivity and specificity from four studies in the meta-analysis phase can be combined with a study-

specific prevalence to obtain PPV and NPV values that are well-calibrated for use in the omitted study. This is a similar idea to Willis and Hyde [156], who suggest examining whether tailored meta-analysis results are needed for individual settings, rather than naively applying summary meta-analysis results everywhere. Note that there is only a small decision tree here to validate using IECV; there are two branches (predicted risk based on either positive or negative PTH values), which is typical in a diagnostic (or short-term prognostic) test situation where a threshold is used to make clinical decisions based on positive (high) or negative (low) levels of a test. However, this situation is similar to the use of a risk score from a prediction model containing multiple variables, which is then dichotomised at a particular threshold (to provide high and low levels) to inform clinical decision making.

Table 5.13 shows the summary sensitivity and specificity values along with their tau-squared values from the univariate meta-analysis for each combination of four studies plus one study excluded, where one study is left out each time and the univariate meta-analysis is performed on the remaining four studies. Also, the prevalence for each of the excluded studies is stated along with the PPV and NPV values.

Table 5.13 shows that the sensitivity and specificity values are similar to those when all five studies are analysed together, although some of these combinations are lower, i.e. when Lam 03 was excluded the sensitivity and specificity were slightly lower whilst the tau-squared values were similar. But when Lombardi 04 was excluded the sensitivity values were higher but the specificity values were similar to the univariate analysis with

all 5 studies. Also, the tau-squared values for sensitivity were much larger at 34.61 for the first three thresholds.

Table 5.13 contains the columns PPV and NPV predicted for the excluded study; these values were calculated in the excluded study using the study's own prevalence, combined with the summary sensitivity and specificity values calculated for the four development studies (akin to approach 5). In Table 5.14, each of the five combinations is shown with E/O having being calculated for each threshold.

Table 5.13: Sensitivity and specificity values when 'internal-external' cross-validation approach compared to the univariate approach for the 1-2 hours' time-point

% PTH Decrease	Sensitivity	Specificity	T^2_{Sens}	T^2_{Spec}	Prevalence in excluded study	PPV predicted for excluded study	NPV predicted for excluded study
Approach 2							
>40	95.56	52.14	0.00	0.00			
>50	95.56	63.31	0.00	0.01			
>60	95.56	80.34	0.00	0.00			
>65	93.33	82.19	0.00	0.02			
>70	93.33	88.03	0.00	0.00			
>80	77.78	94.02	0.00	0.00			
>90	55.56	100.00	0.00	0.06			
Approach 6 Univariate (4 studies) Lam 03 [141] excluded							
>40	93.94	48.35	0.00	0.00	0.31	44.97	94.67
>50	93.94	58.24	0.00	0.00	0.31	50.26	95.53
>60	93.94	76.92	0.00	0.00	0.31	64.65	96.58
>65	90.91	80.22	0.00	0.00	0.31	67.37	95.16
>70	90.91	86.81	0.00	0.00	0.31	75.59	95.51
>80	72.73	92.31	0.00	0.00	0.31	80.95	88.28
>90	47.99	100.00	4.88	0.00	0.31	100.00	81.06
Approach 6 Univariate (4 studies) Lombardi 04 [143] excluded							
>40	99.93	52.66	34.61	0.06	0.31	48.68	99.94
>50	99.93	65.89	34.61	0.04	0.31	56.83	99.95
>60	99.93	79.42	34.61	0.05	0.31	68.57	99.96
>65	93.10	80.50	0.00	0.01	0.31	68.20	96.29
>70	93.10	87.80	0.00	0.00	0.31	77.42	96.59
>80	79.31	94.22	0.00	0.11	0.31	86.04	91.02
>90	44.14	100.00	5.32	0.00	0.31	100.00	79.94
Approach 6 Univariate (4 studies) McLeod 06 [144] excluded							
>40	93.94	56.52	0.00	0.00	0.23	39.22	96.90
>50	93.94	65.45	0.00	0.03	0.23	44.82	97.31
>60	93.94	82.61	0.00	0.00	0.23	61.74	97.86
>65	93.94	84.78	0.00	0.02	0.23	64.83	97.91
>70	93.94	90.22	0.00	0.00	0.23	74.15	98.03
>80	78.79	95.65	0.00	0.00	0.23	84.40	93.79
>90	50.21	100.00	5.93	0.00	0.23	100.00	87.05
Approach 6 Univariate (4 studies) Warren 02 [145] excluded							
>40	97.67	51.37	0.00	0.01	0.18	30.60	99.01

% PTH Decrease	Sensitivity	Specificity	T^2_{Sens}	T^2_{Spec}	Prevalence in excluded study	PPV predicted for excluded study	NPV predicted for excluded study
>50	97.67	63.45	0.00	0.02	0.18	36.97	99.20
>60	97.67	80.73	0.00	0.00	0.18	52.66	99.37
>65	95.35	83.55	0.00	0.04	0.18	55.99	98.79
>70	95.35	87.16	0.00	0.00	0.18	61.98	98.84
>80	79.07	93.58	0.00	0.00	0.18	73.00	95.32
>90	56.52	100.00	3.74	254523.70	0.18	100.00	91.29
Approach 6	Univariate (4 studies) Warren 04 [146] excluded						
>40	95.24	52.29	0.00	0.04	0.11	19.79	98.89
>50	95.24	64.16	0.00	0.05	0.11	24.72	99.09
>60	95.24	81.91	0.00	0.00	0.11	39.42	99.29
>65	92.86	85.11	0.00	0.00	0.11	43.53	98.97
>70	92.86	88.30	0.00	0.00	0.11	49.52	99.01
>80	78.57	94.80	0.00	0.05	0.11	65.13	97.28
>90	56.98	100.00	2.90	2450.74	0.11	100.00	94.95

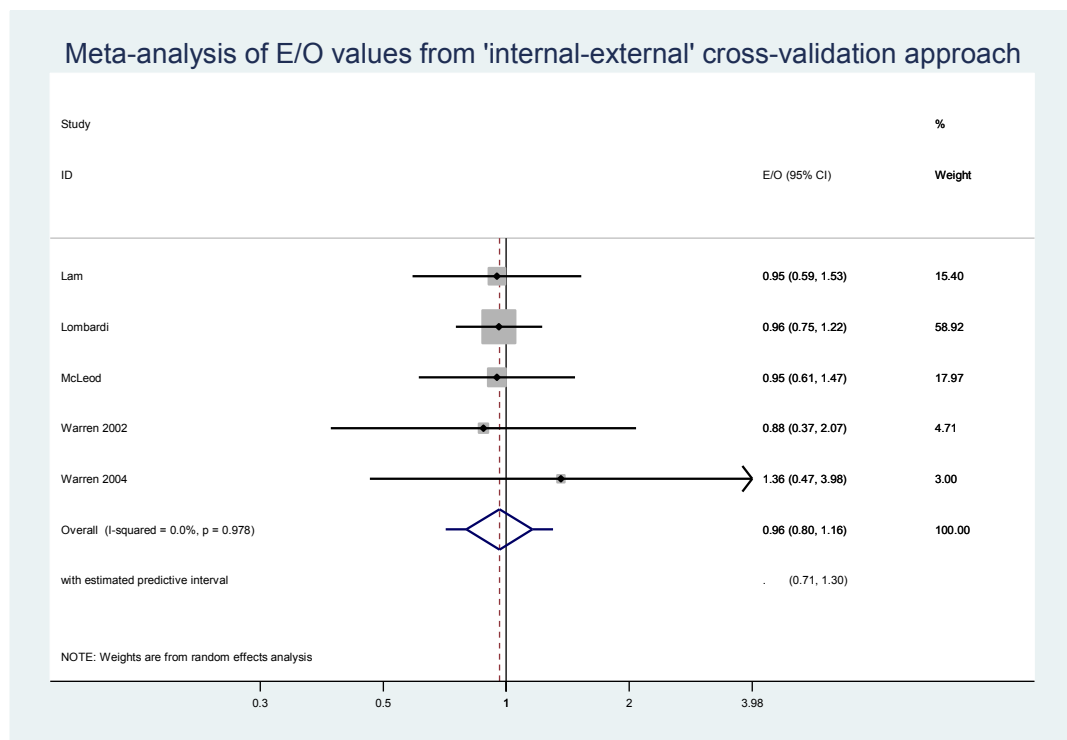
Table 5.14: E/O ratios for all combinations of excluded studies

Threshold	No. of hypocalcaemia who tested positive			95% CI		No. of hypocalcaemia who tested negative			95% CI		Total with hypocalcaemia			95% CI	
	E	O	E/O	Lower	Upper	E	O	E/O	Lower	Upper	E	O	E/O	Lower	Upper
E/O for Lam study [141] when it was excluded															
>40	10.98	12	0.92	0.52	1.61	0.28	0				11.26	12	0.94	0.53	1.65
>50	10.31	12	0.86	0.49	1.51	0.38	0				10.69	12	0.89	0.51	1.57
>60	11.73	12	0.98	0.56	1.72	0.39	0				12.12	12	1.01	0.57	1.78
>65	11.02	11	1.00	0.55	1.81	0.42	1	0.42	0.06	2.97	11.44	12	0.95	0.54	1.68
>70	11.86	11	1.08	0.60	1.95	0.41	1	0.41	0.06	2.94	12.28	12	1.02	0.58	1.80
>80	10.13	9	1.13	0.59	2.16	1.55	3	0.52	0.17	1.60	11.68	12	0.97	0.55	1.71
>90	7.00	7	1.00	0.48	2.10	3.89	5	0.78	0.32	1.87	10.89	12	0.91	0.52	1.60
E/O for Lombardi 04 [143] study when it was excluded															
>40	15.58	15	1.04	0.63	1.72	0.01	1	0.01	0.00	0.08	15.59	16	0.97	0.60	1.59
>50	17.05	15	1.14	0.69	1.89	0.01	1	0.01	0.00	0.07	17.06	16	1.07	0.65	1.74
>60	14.40	15	0.96	0.58	1.59	0.01	1	0.01	0.00	0.09	14.41	16	0.90	0.55	1.47
>65	14.32	15	0.95	0.58	1.58	1.11	1	1.11	0.16	7.90	15.44	16	0.96	0.59	1.57
>70	14.71	15	0.98	0.59	1.63	1.09	1	1.09	0.15	7.75	15.80	16	0.99	0.61	1.61
>80	12.05	12	1.00	0.57	1.77	3.32	4	0.83	0.31	2.21	15.37	16	0.96	0.59	1.57
>90	10.00	10	1.00	0.54	1.86	8.22	6	1.37	0.62	3.05	18.22	16	1.14	0.70	1.86
E/O for McLeod [144] study when it was excluded															
>40	10.98	12	0.92	0.52	1.61	0.28	0				11.26	12	0.94	0.53	1.65
>50	10.31	12	0.86	0.49	1.51	0.38	0				10.69	12	0.89	0.51	1.57
>60	11.73	12	0.98	0.56	1.72	0.39	0				12.12	12	1.01	0.57	1.78
>65	11.02	11	1.00	0.55	1.81	0.42	1	0.42	0.06	2.97	11.44	12	0.95	0.54	1.68
>70	11.86	11	1.08	0.60	1.95	0.41	1	0.41	0.06	2.94	12.28	12	1.02	0.58	1.80
>80	10.13	9	1.13	0.59	2.16	1.55	3	0.52	0.17	1.60	11.68	12	0.97	0.55	1.71
>90	7.00	7	1.00	0.48	2.10	3.89	5	0.78	0.32	1.87	10.89	12	0.91	0.52	1.60
E/O for Warren 02 [145] study when excluded															
>40	1.22	1	1.22	0.17	8.69	0.06	1	0.06	0.01	0.42	1.28	2	0.64	0.16	2.57
>50	1.48	1	1.48	0.21	10.50	0.05	1	0.05	0.01	0.34	1.53	2	0.76	0.19	3.05
>60	1.58	1	1.58	0.22	11.22	0.04	1	0.04	0.01	0.31	1.62	2	0.81	0.20	3.25
>65	1.68	1	1.68	0.24	11.92	0.08	1	0.08	0.01	0.60	1.76	2	0.88	0.22	3.53
>70	0.62	1	0.62	0.09	4.40	0.10	1	0.10	0.01	0.74	0.72	2	0.36	0.09	1.45
>80	0.73	1	0.73	0.10	5.18	0.42	1	0.42	0.06	2.99	1.15	2	0.58	0.14	2.30
>90	0.00	0				0.87	2	0.44	0.11	1.74	0.87	2	0.44	0.11	1.74
E/O for Warren 04 [146] study when excluded															
>40	2.77	3	0.92	0.30	2.86	0.13	0				2.90	3	0.97	0.31	3.00
>50	2.97	3	0.99	0.32	3.07	0.13	0				3.09	3	1.03	0.33	3.20
>60	3.55	3	1.18	0.38	3.67	0.12	0				3.67	3	1.22	0.39	3.79
>65	3.92	3	1.31	0.42	4.05	0.18	0				4.09	3	1.36	0.44	4.23
>70	2.97	3	0.99	0.32	3.07	0.20	0				3.17	3	1.06	0.34	3.28
>80	2.61	2	1.30	0.33	5.21	0.60	1	0.60	0.08	4.25	3.20	3	1.07	0.34	3.31
>90	0.00	0				1.31	3	0.44	0.14	1.36	1.31	3	0.44	0.14	1.36

In Figure 5.3, the E/O values for the 65% threshold are meta-analysed to look at the predictive performance of each study in a meta-analysis. Each of these studies is taken from the IECV approach, where each study was the validation study (excluded from meta-analysis phase). On average it appears the calibration performance of derived PPV and NPV for the predictive test (E/O) is close to 1, with Warren 02 [145] study being 0.88 and Warren 04 [146] study being 1.36 being the worst. However, these are smaller studies as shown by their weighting and actually there is no between-study heterogeneity present as $I^2=0\%$ with a $p=0.98$.

As shown in Figure 5.2, using approach 5 to derive PPV and NPV yields an overall calibration performance result of 0.94 (0.74, 1.19); however, approach 5 did not examine performance in independent data to the meta-analysis. This is resolved by using the IECV approach (as performance is checked in new data), and reassuringly it gives a similar calibration performance result of 0.96 (0.80, 1.16) with a narrower 95% confidence interval. This again shows the PPV and NPV to use in clinical practice need to be tailored from the meta-analysis by using the study-specific prevalence with the meta-analysis results for sensitivity and specificity. Crucially, this finding was missed by Noordzij et al. [94] in their original publication, who recommended PPV and NPV values based on the unstratified prevalence.

Figure 5.3: Meta-analysis of E/O values from 'internal-external' cross validation approach for threshold 65% and time-point 1-2 hours



5.4 Results II -What is the best threshold?

In this final analysis section, the aim is to illustrate issues in choosing the best threshold for PTH based on the univariate and unstratified meta-analysis results

5.4.1 Illustration of how meta-analysis can change the best threshold decision based on sensitivity and specificity alone

Consider first that sensitivity and specificity are deemed equally important, and so the greatest sum from Youden’s statistic (Equation 4.4) is preferred. To illustrate how the unstratified approach and meta-analysis approach may change inferences about the choice of threshold, consider a meta-analysis of just two of the studies (the same two studies providing data for the 6 hour time point, Lam [141] and Lombardi [143]) at the

1-2 hours' time point. Table 5.15 shows the unstratified and univariate results at each threshold, and Table 5.16 indicates that the unstratified approach identifies a 70% threshold value as optimal but the meta-analysis identifies now suggests *either* a 65% or 70% as optimal.

Table 5.15: Sensitivity and Specificity of PTH assay in predicting postoperative hypocalcaemia for 1-2 hours using the same two studies that provided data for the 6 hours analysis; univariate meta-analysis results

1-2 hours % PTH Decrease	Sensitivity	95% CI		Specificity	95% CI	
		Lower	Upper		Lower	Upper
<i>Approach 1</i>						
>40	96.43	82.29	99.37	57.38	44.90	68.98
>50	96.43	82.29	99.37	67.21	54.72	77.66
>60	96.43	82.29	99.37	86.89	76.20	93.20
>65	96.43	82.29	99.37	90.16	80.16	95.41
>70	96.43	82.29	99.37	91.80	82.21	96.45
>72	89.29	72.80	96.29	91.80	82.21	96.45
>80	82.14	64.41	92.12	96.72	88.81	99.10
>90	64.29	45.83	79.29	100.00	94.08	100.00
<i>Approach 2</i>						
>40	96.43	78.58	99.50	57.38	44.76	69.10
>50	97.56	62.64	99.90	69.04	48.94	83.84
>60	96.43	78.58	99.50	86.88	75.90	93.30
>65	96.43	78.58	99.50	90.16	79.79	95.51
>70	96.43	78.58	99.50	90.16	79.79	95.51
>72	90.63	57.84	98.55	91.82	81.73	96.58
>80	83.11	60.27	94.10	97.29	80.95	99.67
>90	64.29	45.38	79.59	100.00		100.00

Table 5.16: Table for Youden's statistic comparing the two approaches for 1-2 hours, using the two studies that have provided data for the 6 hour time point

1-2 hours % PTH Decrease	Youden's statistic	
	Approach 1	Approach 2
>40	153.81	153.81
>50	163.64	166.60
>60	183.32	183.31
>65	186.59	186.59
>70	188.23	186.59
>72	181.09	182.45
>80	178.86	180.37
>90	164.29	164.29

5.4.2 What is the best threshold?

In the original paper by Noordzij et al. [94], the choice of optimal threshold was based on sum of sensitivity and specificity, where harm and benefit were regarded as being equally important. But the best threshold choice will depend on whether it is more

important to have the least number of false negatives only or it is also important to identify the least number of false positives. Now, in this clinical setting, false positives will be patients who are testing positive but will not become hypocalcaemic so actually will be fine and so stay in hospital longer than necessary. But, a false negative will be a patient who *will* become hypocalcaemic but test negative, and so the patient will wrongly be sent home, then develop hypocalcaemia and be treated late, with potentially troublesome consequences and distress. Thus it could be argued that only patients who are wrongly sent home are a concern, and thus minimises the number of false-negatives is a priority.

A threshold formula is now considered, where the aim is to minimise the output value to determine the optimal threshold as specified in the book “Decision making in health and medicine” [157] by Myriam Hunink et al.:

$$Threshold = \frac{harm}{(harm + benefit)}$$

(Equation 5.9)

Values closest to 1 indicates the preferred threshold and this formula has been multiplied by 100 for this thesis.

5.4.2.1 Harm and benefit I

In this instance, regard harm as (1-NPV) and benefit as NPV. Thus we are only interested in those who test negative (those who test positive are ignored).

When we regard harm and benefit as being *equal* this formula then gives us a threshold of:

$$\text{Threshold} = (1 - \text{NPV}) / ((1 - \text{NPV}) + \text{NPV})$$

$$= (1 - \text{NPV})$$

Thus we are using the largest NPV to determine the best threshold.

Table 5.17: Largest NPV for 0-20 minutes

Thresholds	Approach 1		Approach 2	
	PPV	NPV	PPV	NPV
>40	31.00	96.26	32.57	96.39
>50	36.59	97.05	37.12	96.99
>60	43.89	94.66	48.91	94.87
>65	45.75	93.53	51.46	93.76
>70	47.22	93.64	52.35	93.81
>72	48.26	93.72	54.62	93.94
>80	56.00	90.91	56.50	91.23
>90	67.20	84.36	59.25	81.45

Table 5.17 shows that 50% is the best threshold to obtain the largest NPV at the time-point of 0-20 minutes, this means that we consider the best threshold to be the one that has the lowest number of false negatives. False negatives are patients who are incorrectly classified as being negative when they actually will be getting the disease in question. From a patients perspective this is most important, thus this method suggests 50% as the best threshold for this time-point. At other time-points Tables 5.18 and 5.19 suggest that 60% or 65% is the best threshold.

Table 5.18: 1-2 hours

Thresholds	Approach 1		Approach 2	
	PPV	NPV	PPV	NPV
>40	34.89	97.77	35.42	97.71
>50	41.11	98.15	41.71	98.11
>60	56.61	98.54	57.18	98.50
>65	57.13	97.84	59.01	97.82
>70	67.67	98.01	68.17	97.96
>72	70.04	96.92	70.42	96.71
>80	76.15	94.00	78.13	93.90
>90	100.00	89.34	100.00	89.12

Table 5.19: 6 hours

Thresholds	Approach 1		Approach 2	
	PPV	NPV	PPV	NPV
>40	41.27	98.51	41.83	98.47
>50	51.31	98.75	51.89	98.72
>60	64.83	98.90	65.38	98.87
>65	74.69	98.96	75.13	98.94
>70	78.03	97.98	78.42	97.93
>72	82.00	97.05	82.71	97.38
>80	80.04	94.28	80.50	94.23
>90	90.30	90.31	90.50	90.10

5.4.2.2 Harm and benefit II

Now we consider hypothetically that harm outweighs benefit by 20: 1, 50:1 and 100:1, i.e. meaning it is 20/50/100 times worse for someone to be sent home incorrectly. For example, for the 20:1 ratio, Equation 5.9 becomes:

$$(1-NPV)/ ((1-NPV) + (1/20*NPV))$$

Table 5.20: Worse to send someone home incorrectly for 0-20 minutes

Thresholds	20:1		50:1		100:1	
	Approach 1	Approach 2	Approach 1	Approach 2	Approach 1	Approach 2
>40	43.73	42.83	66.02	65.19	79.53	78.93
>50	37.81	38.30	60.31	60.81	75.25	75.63
>60	53.01	51.96	73.83	73.00	84.94	84.39
>65	58.05	57.10	77.57	76.89	87.37	86.94
>70	57.60	56.89	77.25	76.74	87.17	86.84
>72	57.27	56.34	77.01	76.33	87.01	86.58
>80	66.66	65.78	83.33	82.78	90.91	90.58
>90	78.76	82.00	90.26	91.93	94.88	95.79

Regard highest threshold value as the best threshold. These are bold.

Table 5.21: Worse to send someone home incorrectly for 1-2 hours

Thresholds	20:1		50:1		100:1	
	Approach 1	Approach 2	Approach 1	Approach 2	Approach 1	Approach 2
>40	31.33	31.91	53.28	53.96	69.52	70.09
>50	27.38	27.81	48.52	49.06	65.34	65.83
>60	22.86	23.35	42.56	43.23	59.70	60.36
>65	30.63	30.83	52.47	52.70	68.82	69.03
>70	28.88	29.40	50.38	51.01	67.00	67.56
>72	38.86	40.49	61.37	62.98	76.06	77.28
>80	56.07	56.51	76.14	76.46	86.46	86.66
>90	70.47	70.94	85.64	85.92	92.27	92.43

Table 5.22: Worse to send someone home incorrectly for 6 hours

Thresholds	20:1		50:1		100:1	
	Approach 1	Approach 2	Approach 1	Approach 2	Approach 1	Approach 2
>40	23.23	23.71	43.06	43.72	60.20	60.84
>50	20.20	20.59	38.76	39.33	55.87	56.46
>60	18.20	18.61	35.74	36.36	52.66	53.33
>65	17.37	17.65	34.45	34.88	51.24	51.72
>70	29.19	29.71	50.76	51.38	67.34	67.88
>72	37.81	34.98	60.31	57.36	75.25	72.90
>80	54.82	55.05	75.21	75.38	85.85	85.96
>90	68.21	68.73	84.29	84.60	91.47	91.66

Tables 5.20-5.22 show that for 0-20 minutes, 1-2 hours and 6 hours this threshold formula gives a different optimal threshold than found previously in the largest NPV approach which is 90%. This is when harm is regarded as (1-NPV) and benefit is regarded as NPV.

5.4.2.3 Harm and benefit III

In this instance regard harm as (1-NPV) + (1-PPV) and benefit as NPV+PPV

When we regard harm and benefit as being equal this formula then gives us a threshold of:

$$\text{Treatment threshold} = ((1-\text{NPV}) + (1-\text{PPV})) / (((1-\text{NPV}) + (1-\text{PPV})) + (\text{NPV}+\text{PPV}))$$

$$= ((1-\text{NPV}) + (1-\text{PPV}))/2$$

This is given in the tables and also the 20:1, 50:1 and 100:1 hypothetical situations.

I.e. for the 20:1 the threshold is = $((1-\text{NPV}) + (1-\text{PPV})) / (((1-\text{NPV}) + (1-\text{PPV})) + (1/20(\text{NPV}+\text{PPV})))$

Table 5.23: Harm and benefit equal and worse to send someone home for 0-20 minutes

Thresholds	Equal		20:1		50:1		100:1	
	Approach 1	Approach 2	Approach 1	Approach 2	Approach 1	Approach 2	Approach 1	Approach 2
>40	36.37	35.52	91.96	91.68	96.62	96.50	98.28	98.22
>50	33.18	32.95	90.85	90.76	96.13	96.09	98.03	98.01
>60	30.73	28.11	89.87	88.66	95.69	95.13	97.80	97.51
>65	30.36	27.39	89.71	88.30	95.61	94.97	97.76	97.42
>70	29.57	26.92	89.36	88.05	95.45	94.85	97.67	97.36
>72	29.01	25.72	89.10	87.38	95.33	94.54	97.61	97.19
>80	26.55	26.14	87.85	87.62	94.76	94.65	97.31	97.25
>90	24.22	29.65	86.47	89.39	94.11	95.47	96.97	97.68

For this method it appears the choice of optimal threshold differs as the optimal threshold in this instance is 40% (see Table 5.23). This shows that when we assign what is more important in terms of harm and benefit and the different choices available, our choice of optimal threshold can vary drastically..

Table 5.24: Harm and benefit equal and worse to send someone home for 1-2 hours

Thresholds	Equal		20:1		50:1		100:1	
	Approach 1	Approach 2	Approach 1	Approach 2	Approach 1	Approach 2	Approach 1	Approach 2
>40	33.67	33.44	91.03	90.95	96.21	96.17	98.07	98.05
>50	30.37	30.09	89.72	89.59	95.62	95.56	97.76	97.73
>60	22.43	22.16	85.25	85.06	93.53	93.44	96.66	96.61
>65	22.52	21.59	85.32	84.63	93.56	93.23	96.67	96.49
>70	17.16	16.94	80.56	80.31	91.20	91.07	95.39	95.32
>72	16.52	16.44	79.83	79.73	90.82	90.77	95.19	95.16
>80	14.93	13.99	77.82	76.48	89.77	89.05	94.61	94.21
>90	5.33	5.44	52.96	53.50	73.79	74.20	84.92	85.19

Table 5.25: Harm and benefit equal and worse to send someone home for 6 hours

Thresholds	Equal		20:1		50:1		100:1	
	Approach 1	Approach 2	Approach 1	Approach 2	Approach 1	Approach 2	Approach 1	Approach 2
>40	30.11	29.85	89.60	89.49	95.56	95.51	97.73	97.70
>50	24.97	24.70	86.94	86.77	94.33	94.25	97.08	97.04
>60	18.14	17.88	81.59	81.32	91.72	91.58	95.68	95.61
>65	13.18	12.97	75.22	74.87	88.35	88.16	93.82	93.71
>70	12.00	11.83	73.16	72.84	87.20	87.02	93.16	93.06
>72	10.48	9.96	70.06	68.86	85.40	84.68	92.13	91.71
>80	12.84	12.64	74.66	74.31	88.05	87.85	93.64	93.53
>90	9.70	9.70	68.23	68.24	84.30	84.30	91.48	91.48

Tables 5.24-5.25 both give the same optimal threshold of 40%, this is because PPV + NPV and $((1-PPV) + (1-NPV))$ are regarded as having equal importance. What this means is that it is equally important to identify patients who will stay in hospital who will be getting the disease as is identifying the patients who will not get the disease and

who should be sent home. This might be important for a hospital so they can only keep patients who will be getting the disease. But for a patient it is more important to be correctly diagnosed negatively and not to be sent home if they can get the disease. Therefore the largest NPV and the approach where harm and benefit are just regarded as (1-NPV) and NPV respectively is the most important method of defining the optimal threshold from a patient's perspective based on the risks posed by being incorrectly classified as being negative. The original authors say it is important to identify as early as possible those at high risk of hypocalcaemia and decide which patients not to send home. Thus it appears using the NPV (and (1-NPV) for threshold formula) alone is the best way to determine the optimal threshold.

So, the authors' original recommendation of 65% as the best threshold based on the largest combination of sensitivity and specificity in this instance appears reasonable, although in a different scenario the best threshold could have been completely different, as these conclusions were obtained in this section by a different route.

5.5 Discussion

In Chapter 4, the unstratified analysis performed by Noordzij et al. [94] was replicated and then compared to meta-analysis results for *discrimination*. In this chapter, the approaches have now been compared in regards to PPV and NPV, and to assess *calibration* performance using the E/O ratio. This chapter has shown that the method of analysis (unstratified/univariate) can lead to different values for PPV and NPV, as can the choice of prevalence to use in the clinical population for application. For the PTH test, the choice of prevalence was shown to be especially important in improving the

calibration performance for individual studies, so that the PPV and NPV yield an E/O ratio that is closer to 1 in all clinical settings (not just on average) when compared to using the average prevalence. To help establish this, the IECV approach was useful, as it allowed the calibration of PPV and NPV from meta-analysis to be checked in new data (independent to that used to derive the summary sensitivity and specificity results). Even though the analysis to determine the best threshold for PTH agreed with the original author's original recommendation of 65%, this work shows that the value of PPV and NPV for the PTH test depends heavily on the prevalence in the intended population, and so clinicians must use their own population prevalence (but can use the summary sensitivity and specificity meta-analysis results) when deriving PPV and NPV.

5.5.1 Key findings

In Chapter 4, it was found that the sensitivity and specificity results from the three methods (unstratified, univariate and bivariate methods) are often similar, but here the differences are more dramatic when assessing calibration performance of the predictive test. In particular, when using each study's own prevalence the calibration performance of PPV and NPV improves dramatically compared to using the overall unstratified or univariate prevalence.

This chapter has also shown that the 'internal-external' validation approach can be applied to test accuracy research and indicates that this can help to evaluate calibration performance of a test's PPV and NPV values. Here it revealed that, for the PTH test, PPV and NPV need to be derived using the summary sensitivity and specificity results (from meta-analysis), with knowledge of the prevalence in the location of

implementation. This work has further backed up the conclusions obtained by Debray et al. [82] that accounting for heterogeneity in study prevalence is crucial.

Noordzij et al. [94] used their IPD to investigate the predictive ability of PTH for hypocalcaemia. They analysed all the IPD together (unstratified approach) to show its high predictive accuracy. This new meta-analysis further confirmed the ability of PTH assay in predicting hypocalcaemia, but crucially identified that heterogeneity in prevalence can seriously impact upon PTH performance in individual settings, and so population-specific prevalence's must be used when applied the test to new populations. This recommendation was missed by Noordjiz et al. [94], who provided PPV and NPV based on the unstratified prevalence.

Noordzij et al. [94] state in their discussion: "Obtaining a preoperative PTH value is suggested so that percent PTH decrease can be calculated. Routine use of this assay should be considered to improve postoperative management of total and completion thyroidectomy patients. Patients identified as low risk for hypocalcemia could be discharged sooner. Conversely, patients identified as high risk for hypocalcemia developing could be treated earlier, potentially shortening the duration of their hypocalcemic symptoms and hospitalization." With this recommendation in mind, I also performed an analysis to determine the best threshold in terms of minimising the number of false negatives, so that people are not sent home wrongly. In this situation sensitivity and specificity are not equally important, but this had been assumed by Noordzij et al. [94]. Based on this the best threshold appears to be between 60-70% (based on the first two combinations and 1-2 hour and 6-hour time-points). So, the

authors' [94] original recommendation of 65% as the best threshold based on the best combination of sensitivity and specificity in this instance remains reasonable.

Another key finding is that the internal-external cross-validation approach of Royston et al. [101] may be valuable for test accuracy research. This has so far been proposed in a multivariable model context, but this chapter shows that it generalises nicely to the single test/predictor setting and helps examine validation in external data. A paper by Debray et al. [82] uses this internal-external cross-validation process. Debray et al. [82] develop a multivariable logistic regression model from an IPD meta-analysis with potential between-study heterogeneity. They propose strategies for choosing a valid model intercept for when the model is to be validated or applied to new individuals or study samples. This is similar in concept to using a study specific prevalence. Their results indicate that stratified estimation of model intercepts facilitates the derivation of a study specific model intercept, even when it is to be applied to a new study that was not considered during model development. They state [82] that their “framework allows the development (through stratified estimation), implementation in new individuals (through focused intercept choice) and evaluation (through internal-external validation) of a single, integrated prediction model using IPD from multiple studies in order to achieve improved model performance and generalizability.”

5.5.2 What this Chapter adds?

Parts of this chapter have been submitted for publication, and the key findings are summarised in table 5.26. The next part of the thesis is to assess the impact of missing thresholds using a single imputation method, since the full IPD is available for this PTH dataset, this will be used to simulate missing data and to then perform the imputation approach. This is ideal as the full data is already available so it will be straight forward to compare the imputation results to the original data.

Table 5.26: What chapter five adds?

What is known / what is the problem?
<ul style="list-style-type: none">• Chapter four assessed the issue of clustering in this data, and found due to almost zero heterogeneity, the results were similar for the meta-analysis and unstratified approaches• Now the issue is whether we should use these average values for sensitivity and specificity to make prediction in individual settings?• It is important to assess how well this model predicts and compare this to the observed values
What this study adds?
<ul style="list-style-type: none">• The E/O ratio has been calculated using PPV and NPV (which have been calculated using sensitivity, specificity and prevalence)• Initially unstratified prevalence was used but E/O had considerable heterogeneity• When using each study's observed prevalence, E/O had no heterogeneity• If average sensitivity and specificity are combined with each study's observed prevalence, there is reduced error in prediction (almost perfect), and this is new• Currently in Cochrane it is recommended to use the average sensitivity, specificity, which doctors/clinicians may go and use to calculate average PPV and NPV and this results in poor predictions As the PPV and NPV may be wrong if the particular population has a different prevalence to the average• This study shows the average sensitivity and specificity don't give accurate predictions unless combined with each study's observed prevalence
What is needed next?
<ul style="list-style-type: none">• More work is needed in this area, i.e. this example used a dataset that had very little to no heterogeneity within it, if there was heterogeneity this may cause more issues• Can we even use the average sensitivity and specificity? Or will the sensitivity and specificity have to be calculated for each study too?• These are important questions that need to be considered

CHAPTER 6: A SIMULATION STUDY TO EMPIRICALLY EVALUATE AN IMPUTATION METHOD FOR DEALING WITH MISSING THRESHOLDS IN META-ANALYSIS OF A PREDICTIVE TEST

6.1 Introduction

In the evaluation of a potential diagnostic or predictive test, meta-analysis methods are required to synthesise test accuracy from multiple studies. Most meta-analysis methods proposed in the literature use a *single* two by two table from each study, which provides the number of true positives, true negatives, false positives, and false negatives. Such methods univariate and bivariate methods were used in Chapters 3 and 4, applied to each threshold separately. IPD was used in Chapters 4 and 5 to derive the two by two tables, however in practice researchers usually do not have IPD and are reliant on *published* two by two tables. When the test is measured on a continuous scale many published studies report test performance at *multiple* thresholds (often to try and determine the optimal threshold), usually by reporting several two by two tables (one for each threshold) or an ROC curve (to show discriminative ability of each threshold on the same graph). In this situation meta-analysts using published results tend to either utilise the results for just one of the thresholds per study (the most common threshold across studies that is available), or perform a separate meta-analysis for each of the thresholds independently [158]. However, usually the set of thresholds reported by each

study differ, and so the meta-analyst is faced with an incomplete set of threshold results from each study.

The aim of this chapter is to empirically evaluate the use of a linear imputation method, as suggested by Riley et al. [158], for meta-analysis of test accuracy studies when the set of thresholds reported by each study differ. In each study, the method imputes two by two tables for any missing thresholds that are bounded between two reported thresholds; this enables additional studies to be included in each threshold's meta-analysis. In an applied example, Riley et al. [158] found that the method revealed lower diagnostic test accuracy when compared to results from a standard meta-analysis for each threshold independently. This indicates potential selective reporting bias [8, 159], as thresholds are less likely to be reported when they give a lower test accuracy estimate. Riley et al. [158] therefore suggest the method is a useful sensitivity analysis to examine the impact of missing thresholds. However, they have not evaluated their method through any theoretical or simulation procedure.

The aim of this chapter is to assess whether this imputation method works in a case study where IPD are available, and thus all thresholds are known, but through simulation some thresholds are removed to generate 'missing' thresholds. The aims are to evaluate whether the method produces estimates similar to original data, and to examine how the method impact upon the precision and heterogeneity of estimates. Different patterns of selective reporting (different thresholds will be made missing) will be assessed to determine to what degree this method corrects the bias that is introduced. Firstly, the details of the imputation method are described in full, with an applied example, and then the simulation procedure is described and implemented.

The work performed in this chapter has contributed to a publication in *Systematic Reviews* (Riley et al. [43]), where I am the second author. My work has been used to demonstrate whether the imputation method proposed by Riley is robust, using the empirical evaluation based on the PTH data described in full in this chapter. I also contributed heavily to drafting relevant sections and revising the published paper.

6.2 Methodology of the imputation approach

6.2.1 Details of the method

Riley et al. [158] proposed the following approach is used in each study to impute two by two tables for any missing thresholds that are bounded between any two reported thresholds.

Step 1: imputation of missing threshold results between two available thresholds

When a certain threshold has missing results (2 by 2 table is missing), and if thresholds above and below are available, then this missing threshold must have sensitivity and specificity values constrained between the sensitivity and specificity values of the two available thresholds. A linear imputation approach is proposed to impute the results for this missing threshold, as follows.

Firstly sensitivity and specificity values are transformed to the logit scale:

$$\text{Logit- sensitivity} = \ln \left(\frac{\text{sensitivity}}{1 - \text{sensitivity}} \right)$$

(Equation 6.1)

$$\text{Logit- specificity} = \ln \left(\frac{\text{specificity}}{1 - \text{specificity}} \right)$$

(Equation 6.2)

Then by assuming a 1-unit increase in threshold value corresponds to a constant reduction in logit-sensitivity and constant increase in logit-specificity, imputation of the missing logit-sensitivity and logit-specificity are carried out. An example scenario of how this imputation occurs can be shown using the Noordzij et al. [94] thresholds; if thresholds 40% and 65% are available and 50% and 60% are missing (a difference of 25% between the lower and upper thresholds, with 50% being at 2/5 of the difference and 60% being at 4/5 of the difference), then the imputed logit-sensitivity at threshold 50% is (where logit-sens is logit-sensitivity):

$$\text{logit-sens}_{50\%} = \text{logit-sens}_{40\%} + \left(\frac{2(\text{logit-sens}_{65\%} - \text{logit-sens}_{40\%})}{5} \right)$$

(Equation 6.3)

And similarly for 60% the imputed logit-sensitivity is:

$$\text{logit-sens}_{60\%} = \text{logit-sens}_{40\%} + \left(\frac{4(\text{logit-sens}_{65\%} - \text{logit-sens}_{40\%})}{5} \right)$$

(Equation 6.4)

The imputed sensitivity and specificity can be obtained from the imputed logit-sensitivity and logit-specificity by back transforming:

$$\text{imputed sensitivity} = \frac{\exp(\text{imputed logit-sensitivity})}{(1 + \exp(\text{imputed logit-sensitivity}))}$$

(Equation 6.5)

$$\text{imputed specificity} = \frac{\exp(\text{imputed logit-specificity})}{(1 + \exp(\text{imputed logit-specificity}))}$$

(Equation 6.6)

The 2 by 2 table can then be calculated by:

$$\text{imputed true positive} = \text{imputed sensitivity} * \text{total diseased}$$

(Equation 6.7)

$$\text{imputed false negative} = \text{total diseased} - \text{imputed true positive}$$

(Equation 6.8)

$$\text{imputed true negative} = \text{imputed specificity} * \text{total non-diseased}$$

(Equation 6.9)

$$\text{imputed false positive} = \text{total non-diseased} - \text{imputed true negative}$$

(Equation 6.10)

For each pair of observed threshold results, there is a key assumption being made that there is a constant change in the logit values for each 1-unit change in the threshold results. It is important to note that this method does not impute above the highest threshold available or below the lowest threshold available; this requires further assumptions. Also, no imputation is possible if only one threshold was available for a study. If a study has a zero cell in the 2 by 2 table for a threshold then logit-values are not calculable. Thus, in order to impute values on the logit scale as desired, it was necessary to apply a continuity correction by adding 0.5 to each cell in the 2 by 2 table, to ensure that both sensitivity and specificity can be calculated.

Step 2: Meta-analysis at each threshold separately using actual and imputed data

After step 1 has been performed, each threshold has observed and imputed study results, which can be used in a meta-analysis. For example, the bivariate meta-analysis of Chu and Cole [153] at each threshold separately can be applied to the observed and the imputed data. By applying the exact binomial within-study distribution, we avoid the need to apply a continuity correction. This also accounts for any between-study

correlation in sensitivity and specificity. The models were given previously in Chapter 4, but are written here again for ease:

A univariate meta-analysis of sensitivity is:

$$r_{+ve_i} \sim \text{Bin}(\rho_{\text{Sens}_i}, N_{\text{diseased}_i})$$

$$\text{Logit}(\rho_{\text{Sens}_i}) \sim N(\theta_{\text{Sens}}, \tau_{\text{Sens}}^2)$$

(Equation 6.11)

Where r_{+ve_i} is the number of true positives, ρ_{Sens_i} is the true sensitivity in study i , N_{diseased_i} represents the total number with hypocalcaemia, θ_{Sens} is the average logit-sensitivity across studies, τ_{Sens}^2 is the between-study heterogeneity in the logit-sensitivity.

For specificity:

$$r_{-ve_i} \sim \text{Bin}(\rho_{\text{Spec}_i}, N_{\text{non-diseased}_i})$$

$$\text{Logit}(\rho_{\text{Spec}_i}) \sim N(\theta_{\text{Spec}}, \tau_{\text{Spec}}^2)$$

(Equation 6.12)

Where r_{-ve_i} is the number of true negatives, ρ_{Spec_i} is the true specificity in study i , $N_{\text{non-diseased}_i}$ represents the total without hypocalcaemia, θ_{Spec} is the average logit-specificity across studies, τ_{Spec}^2 is the between-study heterogeneity in the logit specificity.

The meta-analysis models utilise the exact binomial distribution within each study, and estimate the average sensitivity and specificity across studies [150, 151], and the variability of sensitivity and specificity across studies.

$$\begin{aligned}
r_{+ve_i} &\sim \text{Bin}(\rho_{\text{Sens}_i}, N_{\text{diseased}_i}) \\
r_{-ve_i} &\sim \text{Bin}(\rho_{\text{Spec}_i}, N_{\text{non-diseased}_i}) \\
\begin{matrix} \text{Logit}(\rho_{\text{Sens}_i}) \\ \text{Logit}(\rho_{\text{Spec}_i}) \end{matrix} &\sim N\left(\begin{matrix} \theta_{\text{Sens}} \\ \theta_{\text{Spec}} \end{matrix}, \begin{bmatrix} \tau_{\text{Sens}}^2 & \tau_{\text{Sens,Spec}} \\ \tau_{\text{Sens,Spec}} & \tau_{\text{Spec}}^2 \end{bmatrix}\right)
\end{aligned}$$

(Equation 6.13)

In this model, the parameters are as defined previously, with the addition of $\tau_{\text{Sens,Spec}}$ which is the between-study covariance in logit sensitivity and logit specificity. Covariance might arise because both sensitivity and specificity are being estimated together, so there may be some correlation between them. The between-study covariance matrix is given, which contains the between-study variances and the between-study correlation in logit-sensitivity and logit-specificity. If between-study correlation is zero, the model then reduces to a separate univariate analysis for each of sensitivity and specificity (equation 6.12). But, this is often poorly estimated at +1 or -1 [160] and therefore it is sensible to adopt two separate univariate models here [159].

The models presented can be estimated using adaptive Gaussian quadrature [161], for example using ‘PROC NLMIXED’ in SAS [162], or the ‘XTMELOGIT’ command in Stata [149]. The Stata [149] code used for meta-analysis has been provided in APPENDIX C.

6.2.2 Example of how this method works for the McLeod study [144]

To illustrate how this methodology works, the McLeod study [144] from the Noordzij et al. [94] dataset will be used. The 65% threshold will be deleted and then imputed using the method of Riley et al.

Figure 6.1: Illustrating the Imputation method using the McLeod study

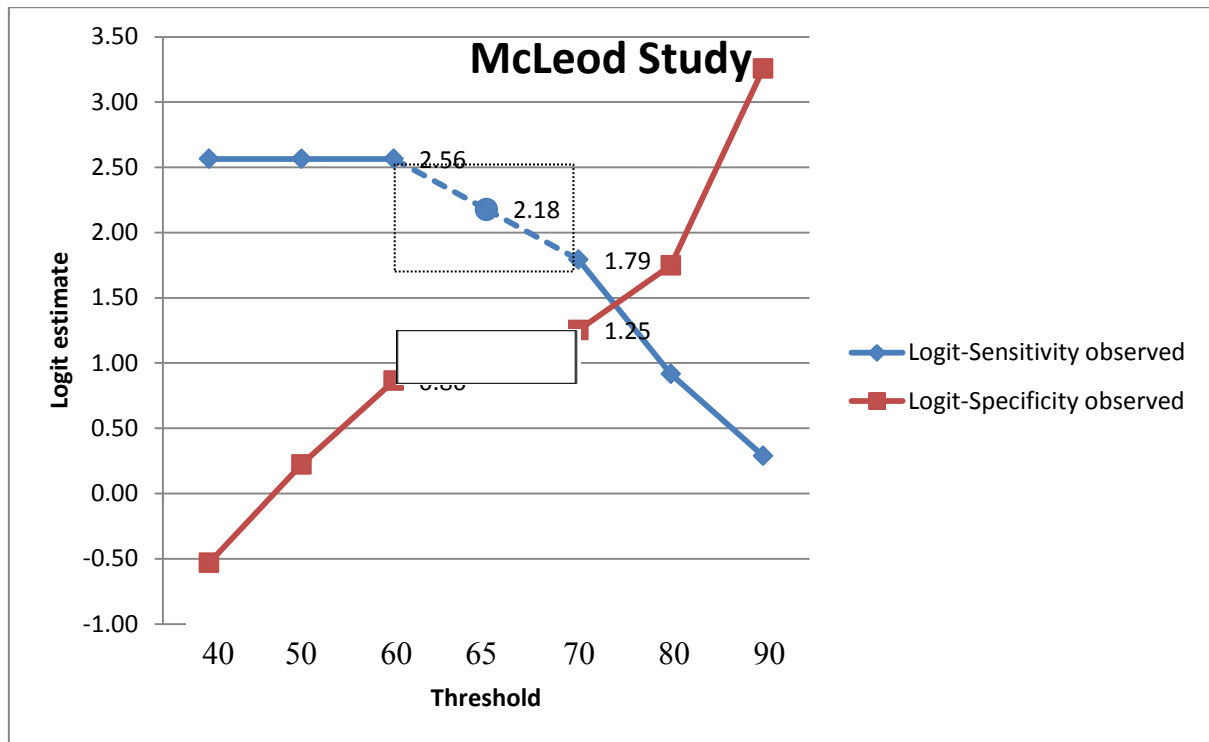


Figure 6.1 illustrates the imputation. Threshold 65% is missing and the imputation method uses the values from 60% and 70% to impute the 65% logit-sensitivity and logit-specificity. A straight dashed line is visible from 60% to 70% with the imputed value being in the middle of this line outlined by a circle. The dashed rectangular boxes around the two imputed values show that this method imputes the threshold at the central point of the region of all logically possible points. In other situations there may be several of these lines for a single study or none at all if only one threshold is present. The imputed 2 by 2 table is presented in Table 6.1; this shows the values for thresholds 60%-70%.

Table 6.1: 2 by 2 table for thresholds 60%-70%, showing what the values are after imputation

Threshold %	TP	FP	FN	TN	Imputed?
60	13	8	1	19	No
65	12.58	6.95	1.42	20.05	Yes
70	12	6	2	21	No

6.2.3 Example of the imputation approach

Consider now a real meta-analysis of a predictive test with missing thresholds.

The clinical question relates to new born babies, and whether the Apgar score is a good test for those babies who will die during the first 28 days of life (known as neonatal mortality) and how the threshold changes sensitivity and specificity.

The word APGAR comes from the acronym (Appearance, Pulse, Grimace, Activity, and Respiration). The Apgar score is determined by evaluating the babies against five criteria which each give a score of 0-2. A low score is worse for a baby, and these scores are usually measured at one minute and five minutes after birth, and repeated later if low scores are observed initially [163]. The Apgar data being used relates to those babies who were < 2500g or pre-term (born before the full term of the pregnancy). Data are available from 11 studies which report up to 10 thresholds, but most only report one or two (Table 6.2). There are 40318 babies in total with 1238 deaths.

Table 6.2: Summarising the studies evaluating the Apgar score

Study name	No of babies	Number of deaths	Outcome Prevalence	Apgar thresholds presented (Sensitivity, Specificity)
Apgar	2422	311	0.128	0 (0.06, 1.00), ≤1 (0.36, 0.96), ≤2 (0.54, 0.93), ≤3 (0.61, 0.90), ≤4 (0.68, 0.85), ≤5 (0.76, 0.79), ≤6 (0.81, 0.71), ≤7 (0.86, 0.59), ≤8 (0.93, 0.35), ≤9 (0.99, 0.10)
Beeby	623	88	0.141	≤3 (0.56, 0.75)
Behnke	748	161	0.215	≤3 (0.70, 0.75), ≤6 (0.92, 0.44)
Drage	1617	151	0.093	≤3 (0.67, 0.90), ≤6 (0.85, 0.71)
Heller	32561	188	0.006	≤3 (0.52, 0.96), ≤6 (0.78, 0.85)
Ikonen	568	100	0.176	≤3 (0.35, 0.95), ≤6 (0.55, 0.85)
Issel	702	72	0.103	≤3 (0.28, 0.93), ≤7 (0.72, 0.73)
Kato	228	6	0.026	≤4 (1.00, 0.66)
Luthy	246	35	0.142	≤3 (0.74, 0.78)
Serenius	211	73	0.346	≤3 (0.38, 0.78)
Tejani	392	53	0.135	≤6 (0.89, 0.58)
Overall	40318	1238	0.031	

Table 6.2 summarises the studies available for meta-analysis and the number of babies, deaths and thresholds in each. The only thresholds that can be imputed are ≤ 4 , ≤ 5 and

≤ 6 , because they are the only thresholds in the studies that have thresholds present that are lower and higher than them, which is essential for this imputation method to work.

Table 6.3: Univariate meta-analysis results for data before and after imputation

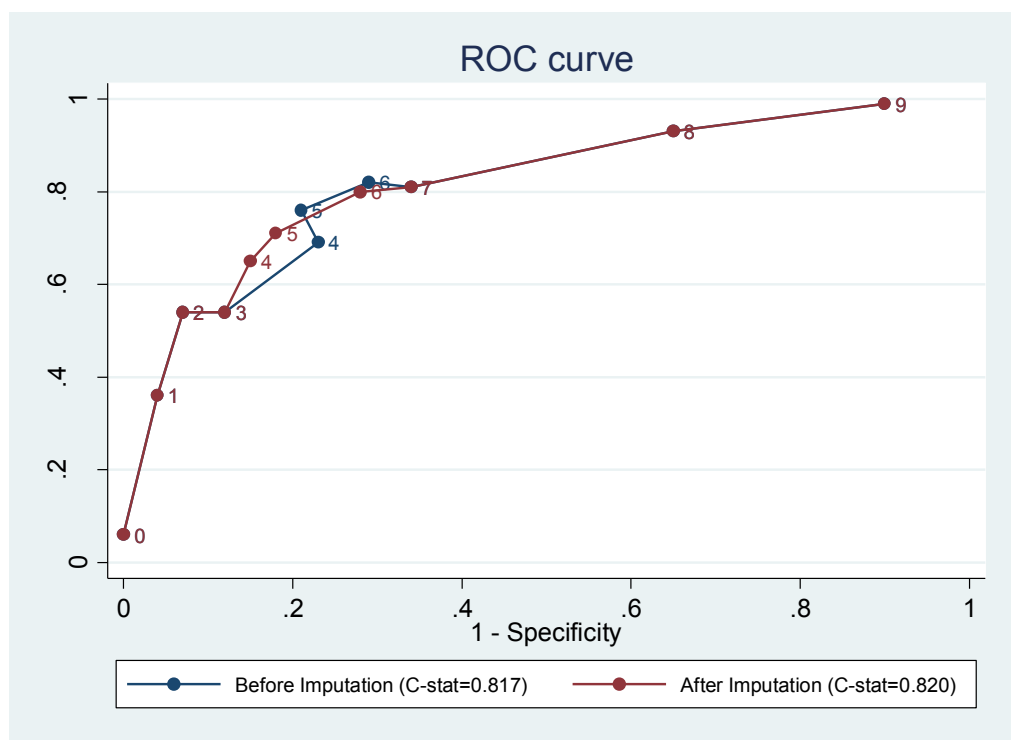
Thresholds	No. of studies contributing data	Summary Sensitivity	T^2_{Sens}	Summary Logit-sens	Standard Error (logit-sens)	Summary Specificity	T^2_{Spec}	Summary Logit-spec	Standard Error (logit-spec)
Non-Imputed									
0	1	0.06	0.00	-2.73	0.06	1.00	0.00	6.27	0.25
≤ 1	1	0.36	0.00	-0.57	0.01	0.96	0.00	3.20	0.01
≤ 2	1	0.54	0.00	0.15	0.01	0.93	0.00	2.52	0.01
≤ 3	9	0.54	0.36	0.14	0.21	0.88	0.57	1.98	0.26
≤ 4	2	0.69	0.00	0.80	0.12	0.77	0.28	1.23	0.38
≤ 5	1	0.76	0.00	1.16	0.02	0.79	0.00	1.34	0.00
≤ 6	6	0.82	0.44	1.51	0.29	0.71	0.49	0.90	0.29
≤ 7	2	0.81	0.12	1.43	0.29	0.66	0.09	0.68	0.22
≤ 8	1	0.93	0.00	2.63	0.05	0.35	0.00	-0.60	0.00
≤ 9	1	0.99	0.00	4.34	0.25	0.10	0.00	-2.23	0.01
Imputed									
≤ 4	7	0.65	0.51	0.60	0.30	0.85	0.55	1.75	0.28
≤ 5	6	0.71	0.47	0.88	0.29	0.82	0.44	1.50	0.27
≤ 6	7	0.80	0.50	1.36	0.28	0.72	0.45	0.97	0.25

Table 6.3 displays the univariate meta-analysis results for each threshold before and after imputation data. For the three points, the sensitivity is slightly lower after imputation and specificity is slightly higher. For example, the ≤ 4 threshold gives a sensitivity of 0.69 and specificity of 0.77 before imputation, and a sensitivity of 0.65 and specificity of 0.85 after imputation. The improvement in specificity is higher than the decrease in sensitivity. This is also indicated by the ROC curves in Figure 6.2, with a slightly higher C statistic of 0.82 after imputation. Note also the smoother, more sensible ROC shape after imputation. Before imputation, the summary specificity wrongly decreased from thresholds 4 to 5, due to the different studies for threshold 4. Interestingly Figure 6.2 displays a non-monotonic summary curve, which is due to the discrepant numbers of studies contributing to the summary points at the thresholds

defining the curve. At threshold 5, only one study contributed data, whereas four studies contributed data at threshold 4 and six studies contributed data to threshold 6, which can explain the potential over optimism in specificity for threshold 5 and the non-monotonic nature of the curve. Now after imputation with more data, the ordering is improved.

In contrast to Riley et al. [158] who showed test accuracy decreased after imputation, these results reveal slightly better test accuracy after imputation. From this meta-analysis, Apgar score appears to be predictive, and a C statistic of 0.82 indicates Apgar score can discriminate rather well between those babies that will have a neonatal death and those that will not.

Figure 6.2: ROC curve comparing before and after imputation data



6.3 Methods for a simulation study of the imputation method

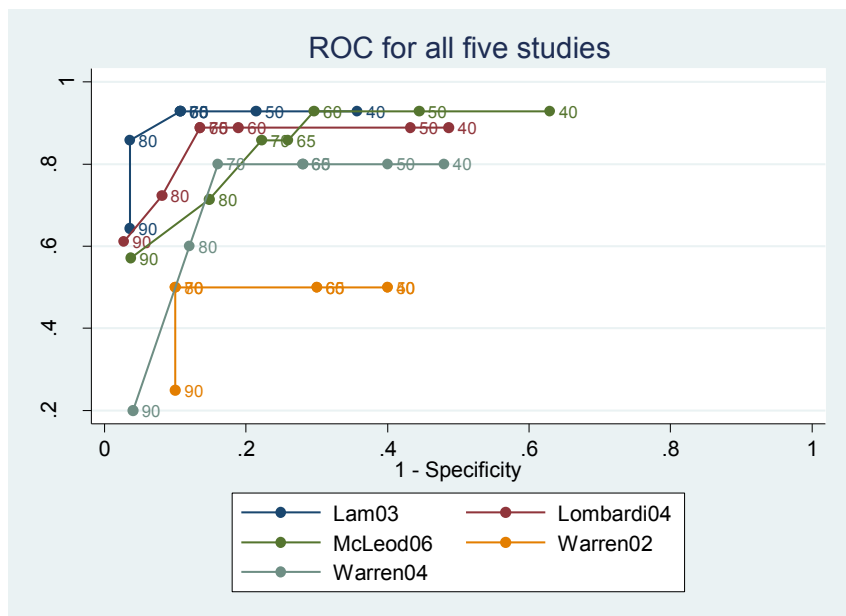
To evaluate the imputation method further, a simulation study is now presented. In this study an existing dataset with all thresholds in all studies is used, so that known meta-

analysis estimates are possible at all thresholds. Then using simulation missing thresholds are generated and the imputation approach applied, to ascertain if it reduces the missing data problem appropriately by 'filling in' the missing data. Four different scenarios of missing thresholds are investigated. The simulation is now detailed.

6.3.1 Dataset

The simulation used the Noordzij dataset [94] introduced previously in Chapter 4, but with a slight modification: a value of one was added to all the 2 by 2 table cells in all studies to ensure no zero cells to ease computational burden. The accuracy of PTH at 1-2 hours is of interest, with five studies available. All thresholds are known in all studies, and hence the true meta-analysis results are available at all thresholds. Seven thresholds are considered: 40%, 50%, 60%, 65%, 70%, 80% and 90%.

Figure 6.3: ROC curves for all five studies for the 1-2 hour time-point from Noordzij data



Results for sensitivity and specificity using PTH at 1-2 hours and 6 hours are summarised in Table 6.4. Most meta-analysis of diagnostic test studies use bivariate

meta-analysis method, as this models the correlation between sensitivity and specificity which is negative in this dataset. There is very little difference between the univariate and bivariate approaches for this dataset, as the between-study heterogeneity for all thresholds is very small or zero and the correlation is poorly estimated when using the bivariate method. This issue was discussed in Chapter 4. In a simulation study the problems of having inestimable correlation structures would cause issues and the univariate method avoids that; thus, given there is no real difference between the estimates from both methods, univariate shall be used from here on in.

Table 6.4: Sensitivity and Specificity of PTH assay in predicting postoperative hypocalcaemia for 1-2 hours: univariate and bivariate results

1-2 hours % PTH Decrease	Sensitivity	95% CI		Specificity	95% CI		T^2_{Sens}	T^2_{Spec}	Correlation
		Lower	Upper		Lower	Upper			
<i>Univariate</i>									
>40	95.6	83.9	98.9	52.1	43.1	61.0	0.0	0.0	-
>50	95.6	83.9	98.9	63.3	53.8	71.9	0.0	0.0	-
>60	95.6	83.9	98.9	80.3	72.2	86.6	0.0	0.0	-
>65	93.3	81.3	97.8	82.2	74.7	88.9	0.0	0.0	-
>70	93.3	81.3	97.8	88.0	80.8	92.8	0.0	0.0	-
>72	88.9	76.0	95.3	89.7	82.8	94.1	0.0	0.0	-
>80	77.8	63.4	87.6	94.0	88.0	97.1	0.0	0.0	-
>90	55.6	41.0	69.2	100.0	Not given	100.0	0.0	0.1	-
<i>Bivariate</i>									
>40	95.6	83.1	99.0	52.2	42.7	61.6	0.1	0.0	-1
>50	95.9	79.6	99.3	63.2	53.0	72.3	0.3	0.0	1
>60	95.6	83.9	98.9	80.3	72.2	86.6	0.0	0.0	0.6
>65	92.9	78.2	98.0	82.6	71.4	90.0	0.2	0.1	1
>70	93.3	81.3	97.8	88.0	80.8	92.8	0.0	0.0	1
>72	88.9	76.0	95.3	89.7	82.8	94.1	0.0	0.0	-1
>80	77.8	63.1	87.8	94.1	87.2	97.4	0.0	0.0	1
>90	55.6	41.0	69.2	100.0	Not given	100.0	0.0	4.7	-0.1

6.3.2 Simulation of 1000 datasets with missing threshold

To assess the performance of the imputation method, missing thresholds were first generated in the Noordzij dataset. Four scenarios of missing data were considered:

- **Scenario I)** Probability of missing data for all thresholds in all studies equals 0.5. This scenario was chosen as it will on average have 50% missing data and reflects those cases where half the results are presented on average.

- **Scenario II)** Thresholds 60% and 70% have a 0.1 probability of being missing, and the remaining thresholds have a 0.5 probability of being missing. This was a selective reporting scenario where most papers would report the most important thresholds which are taken as 60 and 70%. As 65% is the best threshold recommended by Noordzij, this scenario gives the chance to assess how this approach works for the best threshold and also to reflect a more realistic scenario than Scenario I.
- **Scenario III)** Data not missing at random; thresholds with observed sensitivity < 0.90 have a 0.5 probability of being missing, but if observed sensitivity is ≥ 0.90 , then those thresholds are never missing. This is an extreme scenario where data is never missing if sensitivity ≥ 0.90 , assessing how this approach works when certain thresholds are always reported and remaining ones are missing with probability = 0.5.
- **Scenario IV)** A second missing not at random scenario; the first and last thresholds are always present, but data have a 0.5 probability of being missing for the other thresholds if the specificity is < 0.80 . This is an extreme scenario to consider imputation from extreme ends of the threshold spectrum. This reflects an extreme scenario where the first and last thresholds are always presented, and thresholds are missing based on specificity values. This scenario allows all the data to be imputed to assess in a different manner how well it works.

Three steps are then repeated for each scenario.

- **Step 1)** Take the original IPD, and create a new dataset with random functions used to omit data. for example for Scenario I, probability of missing data for all thresholds in all studies equals 0.5.
- **Step 2)** Repeat step 1, until 1000 different meta-analysis datasets are available.
- **Step 3)** Use the imputation method to impute missing threshold results in each study in each of the 1000 datasets.

6.3.3 Meta-analysis of missing data and imputed data

A further two steps are used to meta-analyse the data before and after imputation:

- **Step 4)** Perform a univariate meta-analysis of sensitivity and specificity for each of the 1000 datasets, with and without imputation data, using equation (6.12).
For each meta-analysis, for each threshold the following are calculated: sensitivity, tau-squared for logit-sensitivity, logit-sensitivity, and standard error of logit-sensitivity, specificity, tau-squared for logit-specificity, logit-specificity, and standard error of logit-specificity.
- **Step 5)** Across all 1000 datasets in each scenario, the mean and median of parameter estimates, and coverage of confidence intervals was calculated and compared to the true estimates (from complete data)

Thus in each scenario there are three comparisons of interest: analysis of complete data (the true estimates), analysis with missing data without imputation and analysis with missing data but with imputation. Of interest, is whether meta-analysis results within imputation are preferred to a meta-analysis without imputation, in reference to the original meta-analysis results with complete data.

6.3.4 Programming code

The first step of the programming code involved creating a loop to repeat 1000 times which uses the original dataset. For each repetition, missing data is generated based on each scenario, for example scenario I, the probability of a threshold being missing equals 0.5 for all thresholds in all studies. Once this loop has been run 1000 times, each individual dataset is then appended to create an overall dataset containing the 1000 iterations. This becomes the 'before imputation' dataset, which is meta-analysed and then the Riley method applied to it in order to impute the missing thresholds, before meta-analysis again (see APPENDIX C).

Programming code was developed to perform the meta-analysis. This code took into account that meta-analysis needs to be performed within each dataset and threshold separately, so seven meta-analyses occur within each dataset (seven thresholds), and then this was performed a thousand times (for a thousand datasets). Therefore 14000 (7000 for analysis with data missing and 7000 for analysis with imputed data) meta-analyses were performed for each scenario. A univariate meta-analysis could only be performed for a threshold if two or more studies provided data for it (i.e. they were not missing); if only one study provided data for the threshold then sensitivity and specificity were calculated using the 2 by 2 table, and if no studies provided data then the sensitivity and specificity were not calculated and were missing. The data extracted for each meta-analysis included the pooled logit-sensitivity, standard error of the pooled logit-sensitivity and the tau-squared value for logit-sensitivity, and the equivalent was extracted for logit-specificity.

6.3.5 Statistical measures

To assess bias the pooled estimates for sensitivity and specificity were calculated from the meta-analysis. To assess the precision, the standard errors of the sensitivity and specificity estimates were compared for the different meta-analyses. To assess the heterogeneity of the estimates, the tau-squared estimates were compared. The mean, standard deviation of the mean, median and its upper and lower quartiles were calculated.

6.4 Results of the simulation study of the imputation method

The results are now presented for each simulation scenario separately. The tables give the mean values for the analysis of complete data, along with the mean, standard deviation, median and its upper and lower quartiles for the analysis with data without imputation (contains missing threshold data) and analysis with imputed data (may contain some missing threshold data). These statistical estimates are given for the sensitivity, tau-squared for sensitivity, sensitivity transformed on the logit scale and its standard error, this is repeated for specificity. Box plots are presented to visualise these values, comparing the results for analysis with and without imputation. The standard errors have been plotted against each other from the analysis with data missing versus analysis with imputed data.

6.4.1 Scenario I: Probability of missing equals 0.5 for all thresholds

The summary meta-analysis results, with and without imputation are shown in Table 6.5. Every threshold has on average about 50% missing data before imputation. The results for thresholds 40% and 90% are identical given missing data, as no imputation occurs at these end thresholds. At other thresholds the imputation method reduces the

amount of missing data considerably. For example at threshold 65% the missing data is reduced to just 12.16% of studies after imputation.

6.4.1.1 Pooled estimates

The sensitivity and specificity meta-analysis results before and after imputation are very similar on average and close to the mean estimate from the complete data. For example, for the 65% threshold the mean value from the complete data for sensitivity is 0.85, whilst the mean/median before imputation is 0.84/0.85 and the mean/median after imputation is 0.84/0.84 (see Figures 6.4-6.5).

6.4.1.2 Standard error of pooled estimates

The standard error of the pooled logit-sensitivity and logit-specificity estimates before imputation are much larger than those mean values from the complete data. After imputation the standard errors are much closer to those of true estimated values. For example, at the 65% threshold the mean from the complete data for standard error of logit-sensitivity is 0.38. Before imputation the average standard error is 0.63, but after imputation it is 0.42, much closer to the mean value of 0.38 (see Figures 6.6-6.7).

6.4.1.3 Between-study variances

The τ^2 estimates for the pooled logit-sensitivity and logit-specificity are zero in the true results. The median estimates of τ_{Sens}^2 and τ_{Spec}^2 are also zero before and after imputation.

In summary, scenario I suggests the imputation method obtains estimates on average very close to the estimates from the complete data but with substantially improved precision compared to the analysis with missing data.

Table 6.5: Scenario I, Probability of missing equals 0.5 for all thresholds

Parameter of interest	Mean values from complete data	Meta-analysis without imputation					Meta-analysis with imputed data				
% PTH Decrease		Mean	S.d.	Median	Lower quartile	Upper quartile	Mean	S.d.	Median	Lower quartile	Upper quartile
40%											
*Missing data =49.60%						Missing data =49.60%					
Sensitivity	0.87	0.86	0.08	0.88	0.82	0.90	0.86	0.08	0.88	0.82	0.90
T ² _{Sens}	0.00	0.22	0.38	0.00	0.00	0.21	0.22	0.38	0.00	0.00	0.21
Logit-sens	1.93	1.87	0.51	1.95	1.52	2.22	1.87	0.51	1.95	1.52	2.22
S.E. (logit-sens)	0.40	0.67	0.22	0.61	0.50	0.75	0.67	0.22	0.61	0.50	0.75
Specificity	0.52	0.52	0.06	0.52	0.51	0.56	0.52	0.06	0.52	0.51	0.56
T ² _{Spec}	0.00	0.17	0.36	0.00	0.00	0.06	0.17	0.36	0.00	0.00	0.06
Logit-spec	0.08	0.09	0.24	0.08	0.03	0.24	0.09	0.24	0.08	0.03	0.24
S.E. (logit-spec)	0.18	0.25	0.07	0.24	0.20	0.28	0.25	0.07	0.24	0.20	0.28
50%											
Missing data =49.82%						Missing data =25.46%					
Sensitivity	0.87	0.86	0.08	0.88	0.82	0.90	0.86	0.04	0.87	0.85	0.88
T ² _{Sens}	0.00	0.22	0.38	0.00	0.00	0.21	0.04	0.17	0.00	0.00	0.00
Logit-sens	1.93	1.90	0.49	1.95	1.52	2.22	1.88	0.28	1.93	1.76	1.99
S.E. (logit-sens)	0.40	0.66	0.22	0.61	0.48	0.75	0.50	0.13	0.45	0.41	0.53
Specificity	0.62	0.63	0.05	0.62	0.58	0.66	0.63	0.04	0.63	0.61	0.65
T ² _{Spec}	0.00	0.18	0.36	0.01	0.00	0.07	0.03	0.11	0.00	0.00	0.02
Logit-spec	0.50	0.52	0.24	0.51	0.31	0.65	0.54	0.16	0.55	0.47	0.63
S.E. (logit-spec)	0.18	0.26	0.07	0.26	0.21	0.31	0.23	0.05	0.22	0.19	0.25
60%											
Missing data =50.68%						Missing data =14.64%					
Sensitivity	0.87	0.85	0.08	0.87	0.82	0.89	0.86	0.03	0.85	0.85	0.87
T ² _{Sens}	0.00	0.23	0.38	0.00	0.00	0.21	0.03	0.20	0.00	0.00	0.00
Logit-sens	1.93	1.86	0.51	1.93	1.52	2.14	1.81	0.22	1.77	1.76	1.93
S.E. (logit-sens)	0.40	0.67	0.22	0.61	0.50	0.75	0.44	0.07	0.41	0.39	0.46
Specificity	0.78	0.77	0.05	0.78	0.75	0.81	0.76	0.02	0.76	0.75	0.78
T ² _{Spec}	0.00	0.17	0.36	0.00	0.00	0.07	0.01	0.04	0.00	0.00	0.01
Logit-spec	1.26	1.25	0.28	1.26	1.09	1.42	1.17	0.12	1.17	1.09	1.25
S.E. (logit-spec)	0.21	0.30	0.08	0.30	0.25	0.36	0.24	0.04	0.23	0.21	0.25
65%											
Missing data =51.14%						Missing data =12.16%					
Sensitivity	0.85	0.84	0.07	0.85	0.82	0.88	0.84	0.03	0.84	0.82	0.85
T ² _{Sens}	0.00	0.20	0.38	0.00	0.00	0.11	0.04	0.15	0.00	0.00	0.00
Logit-sens	1.77	1.71	0.45	1.77	1.49	1.98	1.66	0.19	1.64	1.54	1.77
S.E. (logit-sens)	0.38	0.63	0.22	0.56	0.48	0.71	0.42	0.08	0.38	0.37	0.44
Specificity	0.80	0.80	0.05	0.80	0.78	0.83	0.80	0.02	0.80	0.80	0.81
T ² _{Spec}	0.00	0.17	0.37	0.00	0.00	0.02	0.00	0.03	0.00	0.00	0.00
Logit-spec	1.41	1.40	0.30	1.41	1.25	1.58	1.40	0.11	1.41	1.36	1.46
S.E. (logit-spec)	0.22	0.31	0.07	0.31	0.26	0.37	0.24	0.03	0.23	0.22	0.25
70%											
Missing data =48.9%						Missing data =14.74%					
Sensitivity	0.85	0.83	0.08	0.85	0.82	0.88	0.81	0.04	0.82	0.80	0.84
T ² _{Sens}	0.00	0.18	0.36	0.00	0.00	0.00	0.06	0.18	0.00	0.00	0.00
Logit-sens	1.77	1.69	0.48	1.77	1.49	1.98	1.50	0.24	1.50	1.39	1.64
S.E. (logit-sens)	0.38	0.61	0.22	0.53	0.45	0.64	0.41	0.09	0.38	0.35	0.44
Specificity	0.85	0.85	0.03	0.85	0.84	0.87	0.84	0.02	0.84	0.83	0.85

Parameter of interest	Mean values from complete data	Meta-analysis without imputation					Meta-analysis with imputed data				
% PTH Decrease		Mean	S.d.	Median	Lower quartile	Upper quartile	Mean	S.d.	Median	Lower quartile	Upper quartile
T^2_{Spec}	0.00	0.15	0.36	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00
Logit-spec	1.74	1.76	0.20	1.74	1.64	1.90	1.69	0.11	1.68	1.62	1.74
S.E. (logit-spec)	0.25	0.36	0.16	0.33	0.29	0.38	0.27	0.04	0.26	0.24	0.29
80%		Missing data =50.88%					Missing data =26.36%				
Sensitivity	0.73	0.71	0.07	0.73	0.68	0.76	0.70	0.05	0.71	0.68	0.73
T^2_{Sens}	0.00	0.17	0.37	0.00	0.00	0.00	0.04	0.17	0.00	0.00	0.00
Logit-sens	0.98	0.93	0.33	0.98	0.77	1.14	0.87	0.23	0.89	0.77	0.99
S.E. (logit-sens)	0.30	0.45	0.15	0.41	0.35	0.49	0.37	0.10	0.34	0.31	0.41
Specificity	0.91	0.91	0.02	0.91	0.89	0.92	0.91	0.02	0.91	0.90	0.91
T^2_{Spec}	0.00	0.17	0.37	0.00	0.00	0.00	0.02	0.14	0.00	0.00	0.00
Logit-spec	2.26	2.29	0.31	2.27	2.10	2.44	2.27	0.18	2.26	2.17	2.36
S.E. (logit-spec)	0.30	0.46	0.19	0.40	0.36	0.52	0.37	0.08	0.35	0.32	0.39
90%		Missing data =50.76%					Missing data =50.76%				
Sensitivity	0.55	0.51	0.11	0.55	0.50	0.58	0.51	0.11	0.55	0.50	0.58
T^2_{Sens}	0.00	0.20	0.38	0.00	0.00	0.18	0.20	0.38	0.00	0.00	0.18
Logit-sens	0.18	0.05	0.48	0.18	0.00	0.32	0.05	0.48	0.18	0.00	0.32
S.E. (logit-sens)	0.27	0.46	0.26	0.36	0.31	0.47	0.46	0.26	0.36	0.31	0.47
Specificity	0.96	0.96	0.01	0.96	0.96	0.97	0.96	0.01	0.96	0.96	0.97
T^2_{Spec}	0.00	0.18	0.38	0.00	0.00	0.00	0.18	0.38	0.00	0.00	0.00
Logit-spec	3.19	3.16	0.26	3.19	3.07	3.34	3.16	0.26	3.19	3.07	3.34
S.E. (logit-spec)	0.46	0.70	0.18	0.59	0.59	0.73	0.70	0.18	0.59	0.59	0.73

*Average missing data across all 1000 datasets

Figure 6.4: Box plots of Sensitivity, Logit-Sensitivity, S.E.'s for Logit-Sensitivity and Tau-squared for Sensitivity against Threshold %, comparing Non-Imputed and Imputed data (Scenario I)

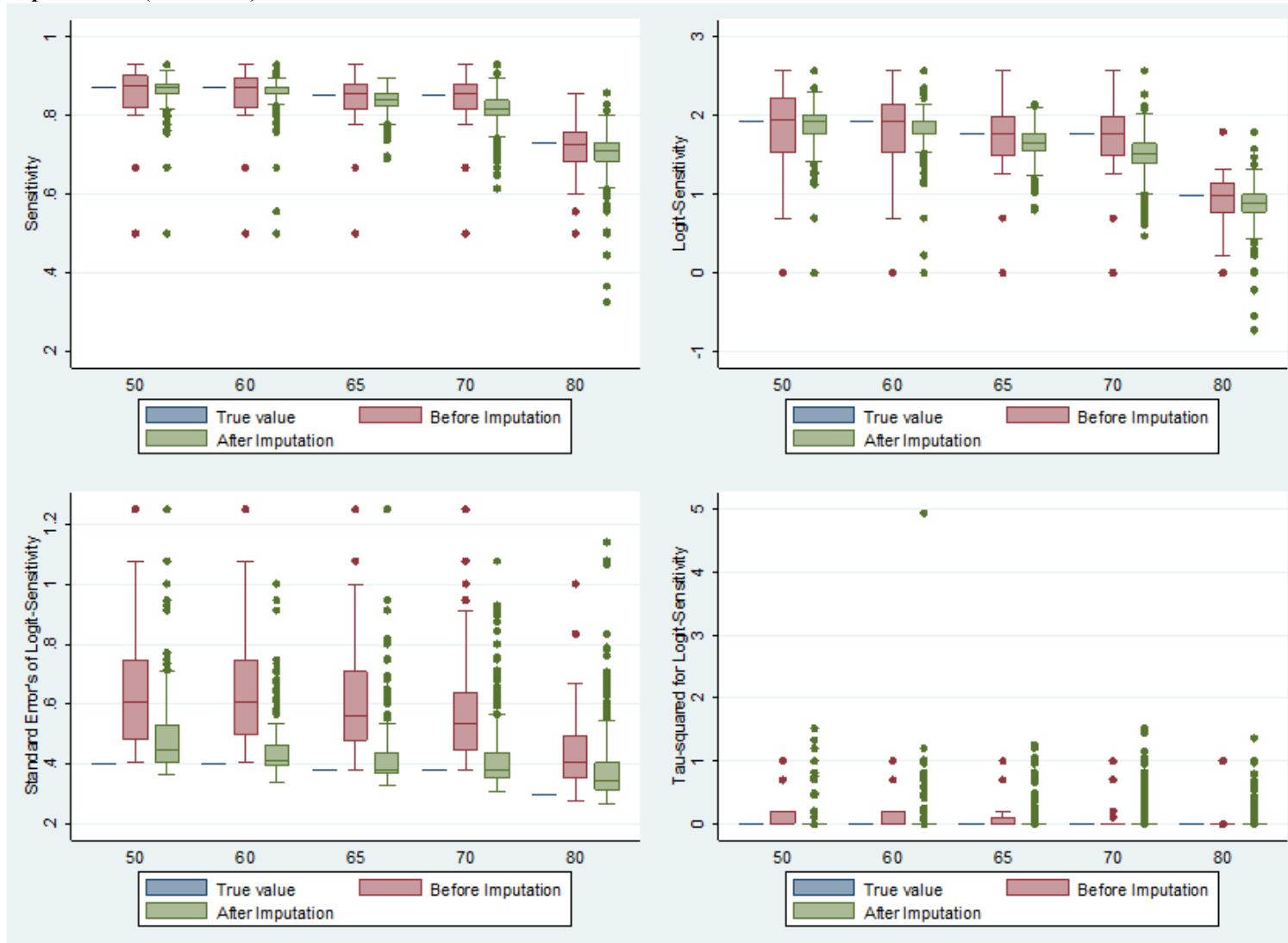


Figure 6.5: Box plots of Specificity, Logit-Specificity, S.E's for Logit-Specificity and Tau-squared for Specificity against Threshold %, comparing Non-Imputed and Imputed data (Scenario I)

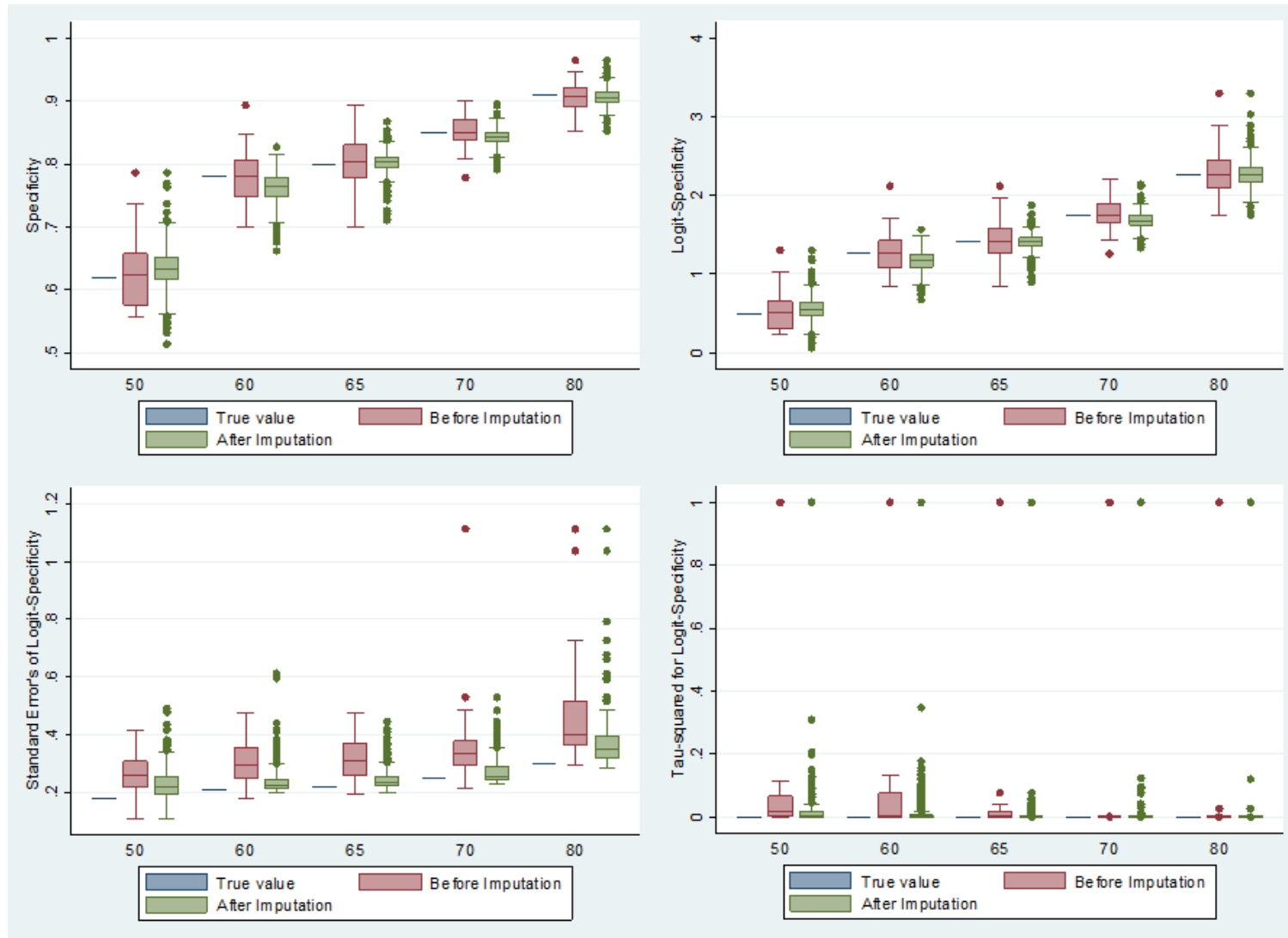


Figure 6.6: S.E's of Logit-Sensitivity for each threshold (Scenario I) in each of the 1000 datasets, before and after imputation

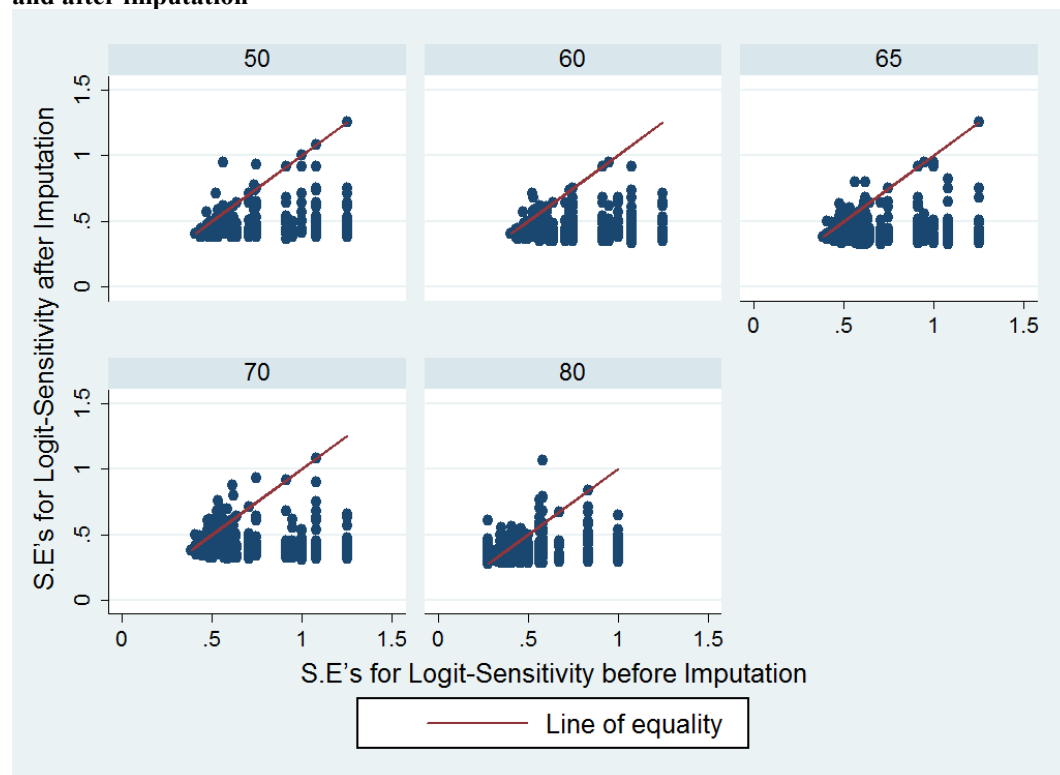


Figure 6.7: S.E's of Logit-Specificity for each threshold (Scenario I) in each of the 1000 datasets, before and after imputation



Although 1000 simulations were performed, there are only 32 different permutations of missing studies for each threshold.

6.4.2 Scenario II: 60% and 70% thresholds have probability missing of 0.1, and the remaining thresholds have probability missing of 0.5

This scenario relates to when some thresholds are more common than others. The results are shown in Table 6.6. Thresholds 60% and 70% have on average 10% of studies missing, whilst the remaining thresholds have on average 50% of studies missing. As before, the imputation method reduces the amount of missing data greatly, especially for threshold 65% with only 2.6% of studies being missing after imputation.

6.4.2.1 Pooled estimates

The average meta-analysis results for sensitivity and specificity are very similar before and after imputation, and when compared to the mean from the complete data. For example, at the 65% threshold the mean from complete data for sensitivity is 0.85. Whilst the mean/median estimate before imputation is 0.83/0.85 and after imputation is 0.85/0.85 (see Figures 6.8-6.9).

6.4.2.2 Standard error of pooled estimates

The average standard errors before imputation are much larger than those from the mean values from complete data, but after imputation the average standard errors are much closer. For example, at the 65% threshold the mean standard error for logit-sensitivity is 0.38, whilst before imputation it is 0.62 on average, but after imputation it is 0.38 on average (see Figures 6.10-6.11).

6.4.2.3 Between-study variances

The τ^2 estimates for pooled logit-sensitivity and logit-specificity are zero in the results from complete data and similar median results are obtained before and after imputation.

Table 6.6: Scenario II, 60% and 70% thresholds have probability missing of 0.1, and the remaining thresholds have probability missing of 0.5

Parameter of interest	Mean values from complete data	Meta-analysis with data missing					Meta-analysis with imputed data				
% PTH Decrease		Mean	S.d.	Median	Lower quartile	Upper quartile	Mean	S.d.	Median	Lower quartile	Upper quartile
50%		Missing data =50.86%					Missing data =24.76%				
Sensitivity	0.87	0.85	0.08	0.87	0.82	0.91	0.87	0.03	0.87	0.85	0.88
T ² _{Sens}	0.00	0.23	0.39	0.00	0.00	0.21	0.04	0.17	0.00	0.00	0.00
Logit-sens	1.93	1.87	0.53	1.93	1.52	2.27	1.91	0.26	1.93	1.76	1.99
S.E. (logit-sens)	0.40	0.67	0.23	0.61	0.48	0.75	0.50	0.13	0.47	0.40	0.52
Specificity	0.62	0.62	0.06	0.62	0.58	0.66	0.63	0.04	0.63	0.61	0.65
T ² _{Spec}	0.00	0.19	0.37	0.01	0.00	0.07	0.03	0.13	0.00	0.00	0.02
Logit-spec	0.50	0.51	0.25	0.50	0.31	0.65	0.53	0.16	0.53	0.47	0.62
S.E. (logit-spec)	0.18	0.26	0.07	0.26	0.21	0.30	0.23	0.04	0.22	0.19	0.25
60%		Missing data =9.88%					Missing data =2.72%				
Sensitivity	0.87	0.87	0.02	0.87	0.87	0.87	0.87	0.01	0.87	0.87	0.87
T ² _{Sens}	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00
Logit-sens	1.93	1.92	0.14	1.93	1.93	1.93	1.91	0.08	1.93	1.93	1.93
S.E. (logit-sens)	0.40	0.43	0.05	0.40	0.40	0.44	0.41	0.02	0.40	0.40	0.40
Specificity	0.78	0.78	0.02	0.78	0.78	0.78	0.78	0.01	0.78	0.77	0.78
T ² _{Spec}	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
Logit-spec	1.26	1.26	0.09	1.26	1.26	1.26	1.25	0.06	1.26	1.22	1.26
S.E. (logit-spec)	0.21	0.23	0.03	0.21	0.21	0.24	0.22	0.01	0.21	0.21	0.21
65%		Missing data =49.72%					Missing data =2.6%				
Sensitivity	0.85	0.83	0.08	0.85	0.82	0.88	0.85	0.01	0.85	0.85	0.85
T ² _{Sens}	0.00	0.20	0.38	0.00	0.00	0.11	0.01	0.06	0.00	0.00	0.00
Logit-sens	1.77	1.69	0.48	1.77	1.49	2.01	1.75	0.08	1.77	1.77	1.77
S.E. (logit-sens)	0.38	0.62	0.21	0.53	0.47	0.64	0.39	0.02	0.38	0.38	0.38
Specificity	0.80	0.80	0.05	0.80	0.78	0.83	0.81	0.01	0.80	0.80	0.81
T ² _{Spec}	0.00	0.17	0.37	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00
Logit-spec	1.41	1.39	0.30	1.41	1.25	1.58	1.42	0.06	1.41	1.41	1.46
S.E. (logit-spec)	0.22	0.31	0.07	0.29	0.26	0.37	0.23	0.01	0.23	0.22	0.23
70%		Missing data =9.42%					Missing data =2.66%				
Sensitivity	0.85	0.85	0.02	0.85	0.85	0.85	0.85	0.01	0.85	0.84	0.85
T ² _{Sens}	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.08	0.00	0.00	0.00
Logit-sens	1.77	1.77	0.12	1.77	1.77	1.77	1.73	0.10	1.77	1.63	1.77
S.E. (logit-sens)	0.38	0.41	0.04	0.38	0.38	0.38	0.39	0.03	0.38	0.38	0.38
Specificity	0.85	0.85	0.01	0.85	0.85	0.85	0.85	0.01	0.85	0.85	0.85
T ² _{Spec}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Logit-spec	1.74	1.74	0.06	1.74	1.74	1.74	1.73	0.05	1.74	1.74	1.74
S.E. (logit-spec)	0.25	0.26	0.03	0.25	0.25	0.28	0.25	0.01	0.25	0.25	0.25
80%		Missing data =51.48%					Missing data =25.54%				
Sensitivity	0.73	0.71	0.07	0.72	0.68	0.76	0.71	0.05	0.71	0.69	0.73
T ² _{Sens}	0.00	0.18	0.39	0.00	0.00	0.00	0.04	0.16	0.00	0.00	0.00
Logit-sens	0.98	0.93	0.35	0.96	0.77	1.14	0.89	0.23	0.89	0.80	1.00
S.E. (logit-sens)	0.30	0.46	0.16	0.41	0.35	0.49	0.37	0.10	0.35	0.31	0.39
Specificity	0.91	0.90	0.03	0.90	0.89	0.92	0.90	0.01	0.91	0.90	0.91

Parameter of interest	Mean values from complete data	Meta-analysis with data missing					Meta-analysis with imputed data				
		Mean	S.d.	Median	Lower quartile	Upper quartile	Mean	S.d.	Median	Lower quartile	Upper quartile
% PTH Decrease											
T^2_{Spec}	0.00	0.18	0.39	0.00	0.00	0.00	0.02	0.12	0.00	0.00	0.00
Logit-spec	2.26	2.29	0.33	2.23	2.08	2.43	2.26	0.17	2.26	2.17	2.35
S.E. (logit-spec)	0.30	0.47	0.20	0.40	0.36	0.52	0.37	0.08	0.35	0.32	0.39

*Average missing data across all 1000 datasets

Figure 6.8: Box plots of Sensitivity, Logit-Sensitivity, S.E's for Logit-Sensitivity and Tau-squared for Sensitivity against Threshold %, comparing Non-Imputed and Imputed data (Scenario II)

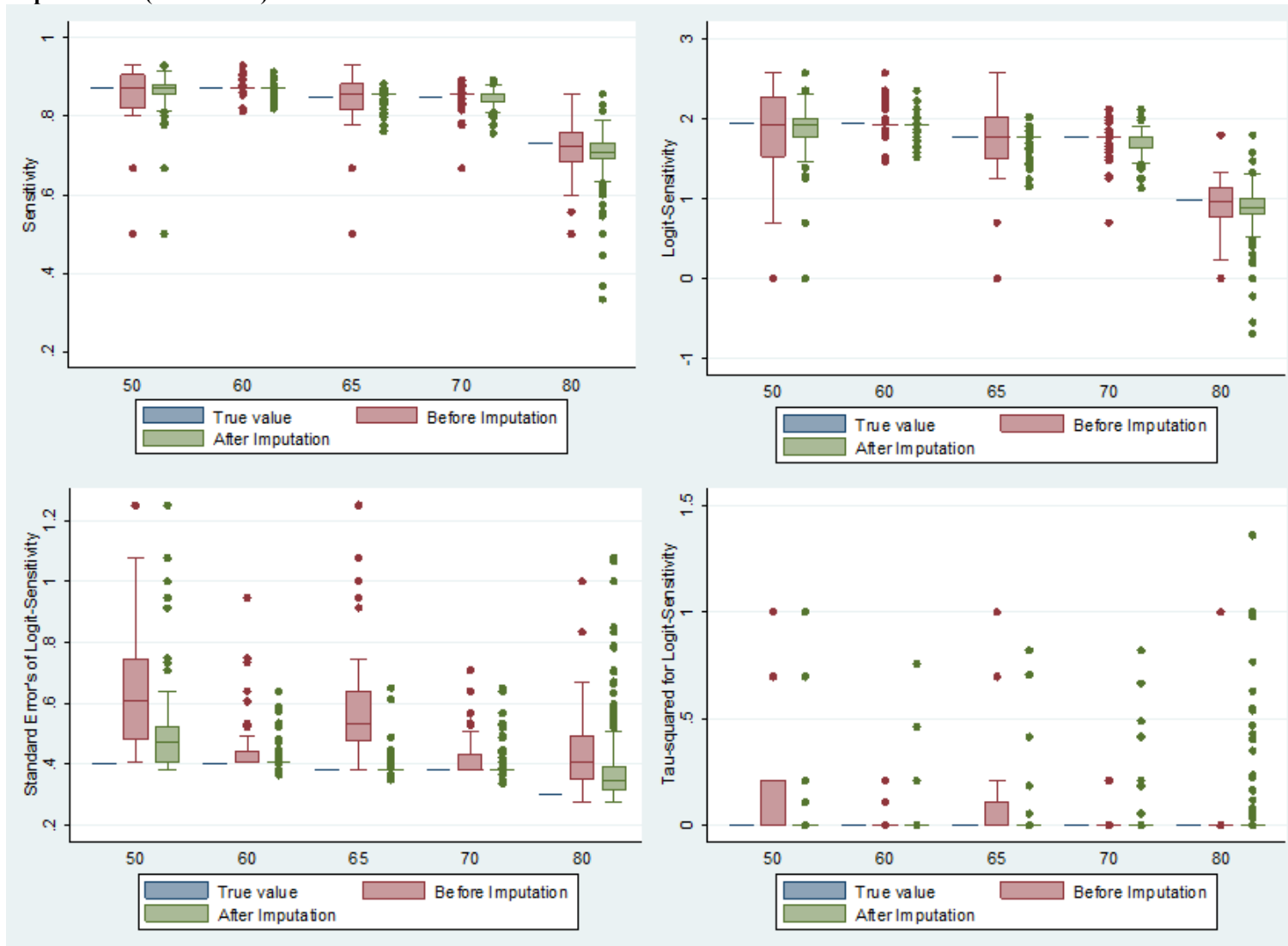


Figure 6.9: Box plots of Specificity, Logit-Specificity, S.E's for Logit-Specificity and Tau-squared for Specificity against Threshold %, comparing Non-Imputed and Imputed data (Scenario II)

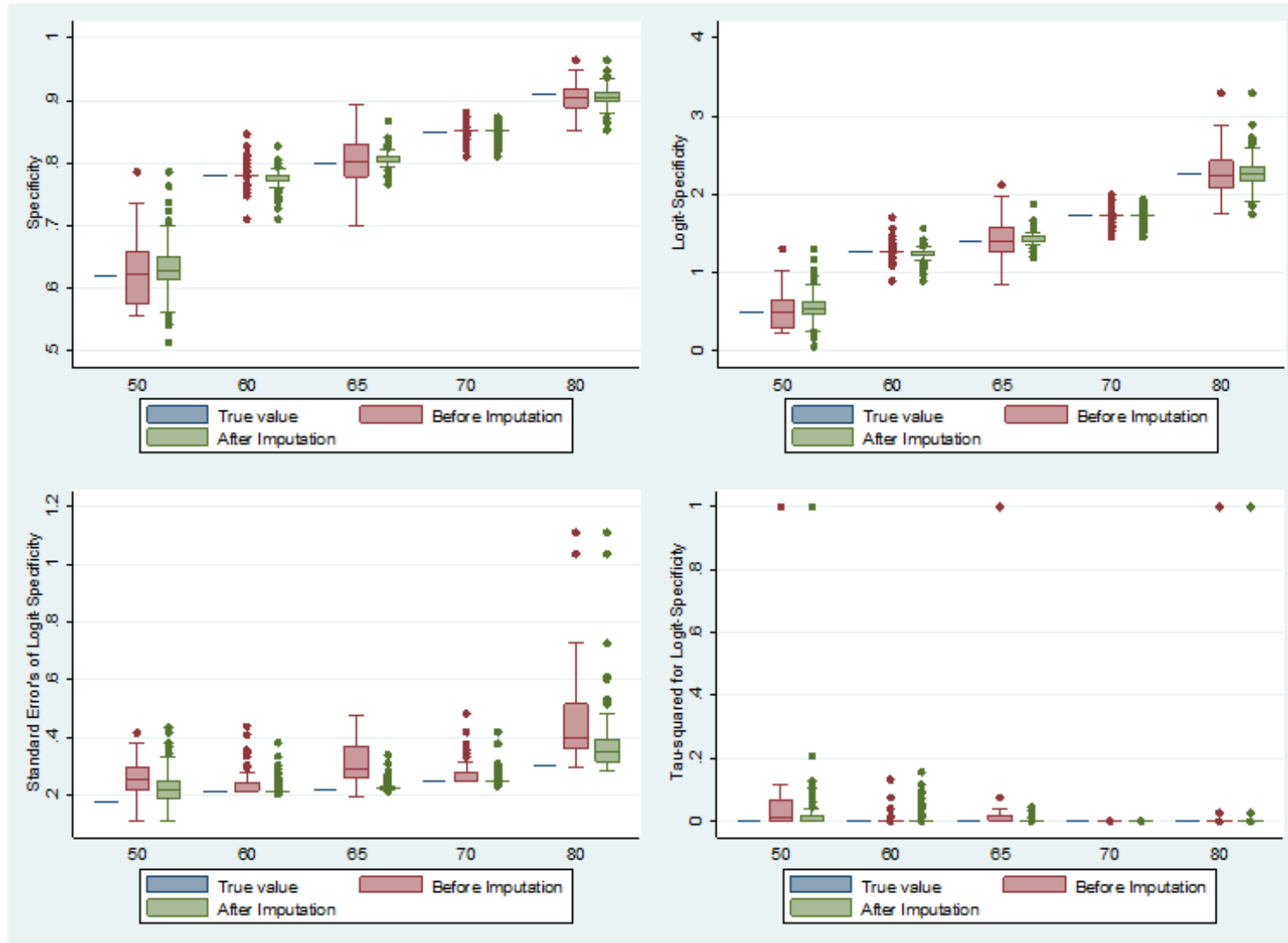


Figure 6.10: S.E's of Logit-Sensitivity for each threshold (Scenario II) in each of the 1000 datasets, before and after imputation

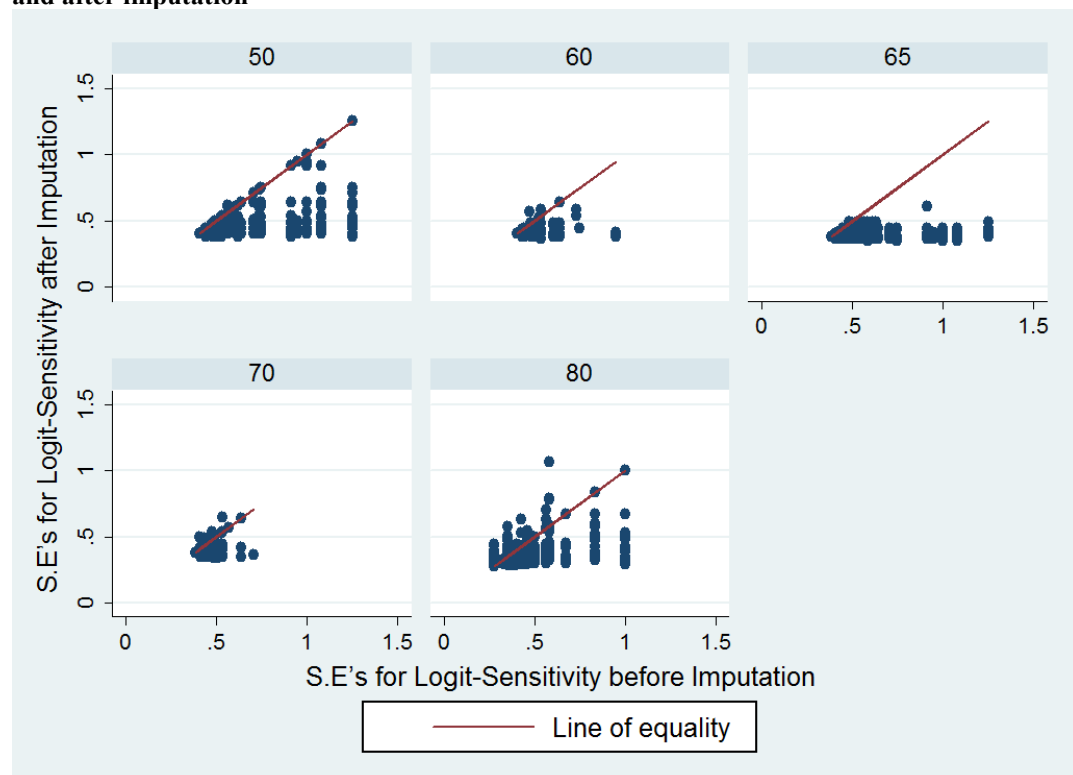


Figure 6.11: S.E's of Logit-Specificity for each threshold (Scenario II) in each of the 1000 datasets, before and after imputation



Although 1000 simulations were performed, there are only 32 different permutations of missing studies for each threshold.

6.4.3 Scenario III: Data not missing at random, thresholds with sensitivity < 0.90 having a probability of 0.5 of being missing, but otherwise are reported

The results of this scenario are shown in Table 6.7. This scenario is a missing not at random scenario, where if sensitivity ≥ 0.90 , it is always reported but not otherwise (selective reporting). This is a realistic scenario, as biased reporting is a major problem in observation research [164]. Missing data is again reduced impressively after imputation. For example, at threshold 65% before imputation 39.88% of studies are missing, but after imputation 8.34% are missing.

6.4.3.1 Pooled estimates

The sensitivity and specificity meta-analysis results before imputation are on average being overestimated in comparison to the complete data values (see Figures 6.12-6.13), due to the selective reporting. However, the use of imputed data reduces this over estimation. For example, at the 50% threshold the mean value from complete data for sensitivity is 0.87, whilst the median estimate before imputation is 0.90 but the median estimate after imputation is 0.87. For the 80% threshold, the mean value from complete data for sensitivity is 0.73, whilst the median estimate before imputation is 0.75 and the median estimate after imputation is 0.71. In the latter, the imputation method underestimates.

6.4.3.2 Standard error of pooled estimates

As in other scenarios, the standard errors before imputation occurs are much larger than the values from the complete data, but after imputation these are closer to the complete data values (see Figures 6.14-6.15). For example, for threshold 50% the mean standard error for the pooled logit-sensitivity is 0.40, whilst the median standard error before imputation is 0.52 and the median after imputation is 0.44.

6.4.3.3 Between-study variances

As previously, the τ^2 estimates are zero in the results from complete data and before and after imputation.

Table 6.7: Scenario III; Data not missing at random, thresholds with sensitivity<0.90 having a probability of 0.5 of being missing, but otherwise are reported

Parameter of interest	Mean values from complete data	Meta-analysis with data missing					Meta-analysis with imputed data				
% PTH Decrease		Mean	S.d.	Median	Lower quartile	Upper quartile	Mean	S.d.	Median	Lower quartile	Upper quartile
50%											
		Missing data =30.58%					Missing data =15.5%				
Sensitivity	0.87	0.89	0.02	0.90	0.88	0.91	0.88	0.02	0.87	0.87	0.88
T ² _{Sens}	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.12	0.00	0.00	0.00
Logit-sens	1.93	2.15	0.24	2.22	1.95	2.35	2.01	0.17	1.93	1.93	2.01
S.E. (logit-sens)	0.40	0.53	0.10	0.52	0.47	0.61	0.46	0.07	0.44	0.40	0.48
Specificity	0.62	0.64	0.02	0.64	0.63	0.66	0.64	0.02	0.64	0.62	0.65
T ² _{Spec}	0.00	0.04	0.04	0.04	0.02	0.07	0.01	0.02	0.00	0.00	0.02
Logit-spec	0.50	0.59	0.08	0.59	0.53	0.67	0.58	0.07	0.59	0.51	0.63
S.E. (logit-spec)	0.18	0.26	0.06	0.26	0.23	0.31	0.21	0.04	0.20	0.18	0.23
60%											
		Missing data =29.64%					Missing data =8.72%				
Sensitivity	0.87	0.89	0.02	0.90	0.87	0.91	0.87	0.02	0.87	0.87	0.87
T ² _{Sens}	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.16	0.00	0.00	0.00
Logit-sens	1.93	2.14	0.23	2.22	1.93	2.30	1.94	0.15	1.93	1.93	1.93
S.E. (logit-sens)	0.40	0.52	0.10	0.48	0.45	0.53	0.44	0.06	0.40	0.40	0.47
Specificity	0.78	0.79	0.01	0.79	0.78	0.79	0.77	0.01	0.77	0.77	0.78
T ² _{Spec}	0.00	0.03	0.04	0.00	0.00	0.04	0.01	0.02	0.00	0.00	0.00
Logit-spec	1.26	1.31	0.08	1.30	1.25	1.35	1.23	0.07	1.22	1.19	1.26
S.E. (logit-spec)	0.21	0.28	0.07	0.26	0.23	0.30	0.23	0.03	0.21	0.21	0.24
65%											
		Missing data =39.88%					Missing data =8.34%				
Sensitivity	0.85	0.87	0.03	0.86	0.84	0.89	0.85	0.02	0.85	0.84	0.85
T ² _{Sens}	0.00	0.06	0.18	0.00	0.00	0.00	0.04	0.15	0.00	0.00	0.00
Logit-sens	1.77	1.89	0.24	1.82	1.69	2.11	1.71	0.16	1.77	1.63	1.77
S.E. (logit-sens)	0.38	0.54	0.14	0.49	0.44	0.61	0.41	0.06	0.38	0.38	0.43
Specificity	0.80	0.82	0.03	0.82	0.80	0.84	0.81	0.01	0.80	0.80	0.81
T ² _{Spec}	0.00	0.01	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Logit-spec	1.41	1.51	0.18	1.51	1.39	1.64	1.43	0.09	1.41	1.36	1.46
S.E. (logit-spec)	0.22	0.31	0.06	0.28	0.26	0.36	0.24	0.02	0.23	0.22	0.25
70%											
		Missing data =39.72%					Missing data =11.38%				
Sensitivity	0.85	0.87	0.03	0.86	0.84	0.89	0.83	0.03	0.84	0.81	0.85
T ² _{Sens}	0.00	0.06	0.18	0.00	0.00	0.00	0.08	0.20	0.00	0.00	0.00
Logit-sens	1.77	1.88	0.23	1.82	1.69	2.11	1.59	0.20	1.63	1.46	1.76
S.E. (logit-sens)	0.38	0.54	0.14	0.49	0.44	0.61	0.41	0.08	0.38	0.36	0.45
Specificity	0.85	0.86	0.02	0.85	0.85	0.87	0.84	0.01	0.84	0.83	0.85
T ² _{Spec}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Logit-spec	1.74	1.79	0.16	1.76	1.70	1.90	1.70	0.10	1.69	1.62	1.74
S.E. (logit-spec)	0.25	0.33	0.07	0.31	0.29	0.38	0.26	0.03	0.25	0.24	0.28
80%											
		Missing data =49.86%					Missing data =25.18%				
Sensitivity	0.73	0.75	0.07	0.75	0.70	0.79	0.71	0.04	0.71	0.69	0.73
T ² _{Sens}	0.00	0.02	0.11	0.00	0.00	0.00	0.05	0.18	0.00	0.00	0.00
Logit-sens	0.98	1.15	0.42	1.07	0.83	1.30	0.89	0.21	0.89	0.78	1.00
S.E. (logit-sens)	0.30	0.45	0.12	0.41	0.36	0.49	0.37	0.10	0.34	0.31	0.39
Specificity	0.91	0.90	0.03	0.90	0.89	0.91	0.91	0.01	0.91	0.90	0.91

Parameter of interest	Mean values from complete data	Meta-analysis with data missing					Meta-analysis with imputed data				
		Mean	S.d.	Median	Lower quartile	Upper quartile	Mean	S.d.	Median	Lower quartile	Upper quartile
% PTH Decrease											
T^2_{Spec}	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.12	0.00	0.00	0.00
Logit-spec	2.26	2.20	0.29	2.23	2.05	2.35	2.28	0.17	2.26	2.18	2.36
S.E. (logit-spec)	0.30	0.41	0.10	0.39	0.35	0.47	0.36	0.07	0.35	0.32	0.39

*Average missing data across all 1000 datasets

Figure 6.12: Box plots of Sensitivity, Logit-Sensitivity, S.E's for Logit-Sensitivity and Tau-squared for Sensitivity against Threshold %, comparing Non-Imputed and Imputed data (Scenario III)

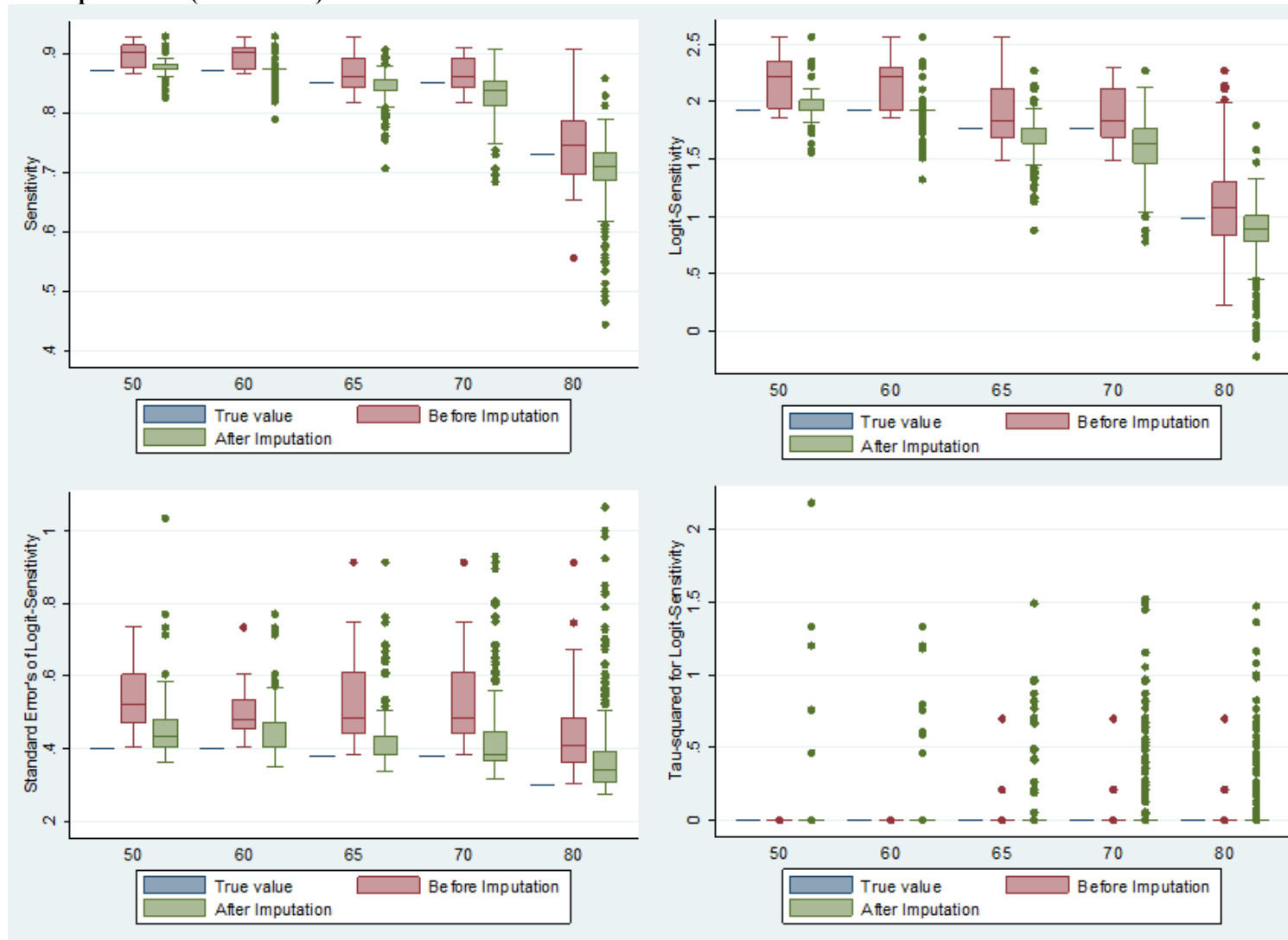


Figure 6.13: Box plots of Specificity, Logit-Specificity, S.E's for Logit-Specificity and Tau-squared for Specificity against Threshold %, comparing Non-Imputed and Imputed data (Scenario III)

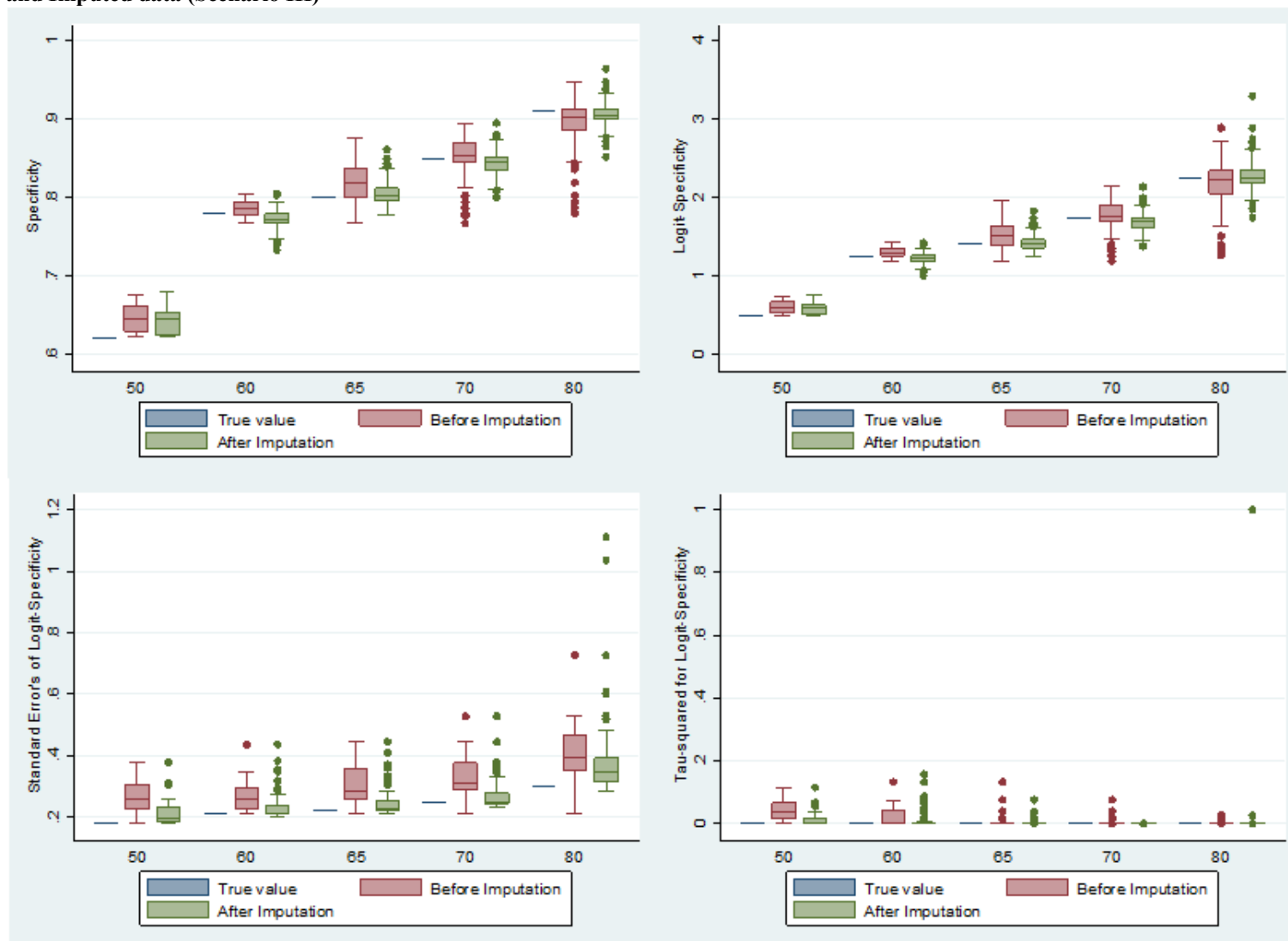


Figure 6.14: S.E's of Logit-Sensitivity for each threshold (Scenario III) in each of the 1000 datasets, before and after imputation



Figure 6.15: S.E's of Logit-Specificity for each threshold (Scenario III) in each of the 1000 datasets, before and after imputation



Although 1000 simulations were performed, there are only 32 different permutations of missing studies for each threshold.

6.4.4 Scenario IV: Missing not at random, with the first and last thresholds always being present. The data has a probability of missing equals 0.5 for the other thresholds if the specificity is <0.80

This case study is a second missing not at random scenario, where the thresholds within each study have a 50% chance of missing if the specificity < 0.80 . But they are not missing for specificity ≥ 0.80 *and* the first and last threshold. This is an extreme 'selective reporting' scenario to allow the imputation method to always have the end thresholds available, thus always enabling missing thresholds to be imputed. Table 6.8 shows the results for this scenario.

The pooled estimates, standard errors and between-study variance show similar results as previous scenarios. In particular, the standard errors are considerably smaller after imputation, and the pooled estimates are close to the mean values from the complete data (either with or without imputation). There is 0% missing data after imputation; this is due to the first and last threshold always being present, resulting in every missing threshold being imputed.

Table 6.8: Scenario IV; missing not at random, with the first and last thresholds always being present. The data has a probability of missing equals 0.5 for the other thresholds if the specificity is <0.80

Parameter of interest	Mean values from complete data	Meta-analysis with data missing					Meta-analysis with imputed data				
% PTH Decrease		Mean	S.d.	Median	Lower quartile	Upper quartile	Mean	S.d.	Median	Lower quartile	Upper quartile
50%		Missing data =49.54%					Missing data =0%				
Sensitivity	0.87	0.86	0.05	0.87	0.85	0.89	0.87	0.01	0.87	0.87	0.87
T ² _{Sens}	0.00	0.06	0.18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Logit-sens	1.93	1.90	0.35	1.93	1.76	2.14	1.89	0.07	1.93	1.93	1.93
S.E. (logit-sens)	0.40	0.57	0.16	0.53	0.44	0.64	0.40	0.01	0.40	0.40	0.40
Specificity	0.62	0.61	0.06	0.60	0.57	0.65	0.64	0.02	0.64	0.62	0.65
T ² _{Spec}	0.00	0.02	0.03	0.00	0.00	0.03	0.01	0.01	0.00	0.00	0.01
Logit-spec	0.50	0.45	0.26	0.41	0.27	0.62	0.56	0.07	0.56	0.50	0.63
S.E. (logit-spec)	0.18	0.25	0.06	0.25	0.20	0.30	0.19	0.01	0.19	0.18	0.19
60%		Missing data =29.62%					Missing data =0%				
Sensitivity	0.87	0.88	0.02	0.88	0.86	0.90	0.86	0.01	0.87	0.85	0.87
T ² _{Sens}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Logit-sens	1.93	2.05	0.20	1.99	1.82	2.22	1.84	0.09	1.93	1.77	1.93
S.E. (logit-sens)	0.40	0.48	0.06	0.48	0.44	0.52	0.39	0.01	0.40	0.38	0.40
Specificity	0.78	0.81	0.02	0.80	0.79	0.81	0.77	0.01	0.77	0.76	0.78
T ² _{Spec}	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.03	0.01	0.00	0.04
Logit-spec	1.26	1.43	0.13	1.41	1.35	1.46	1.22	0.05	1.22	1.18	1.26
S.E. (logit-spec)	0.21	0.26	0.04	0.26	0.24	0.27	0.22	0.01	0.22	0.21	0.23
65%		Missing data =29.36%					Missing data =0%				
Sensitivity	0.85	0.87	0.02	0.86	0.85	0.89	0.85	0.01	0.85	0.85	0.85
T ² _{Sens}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Logit-sens	1.77	1.96	0.18	1.82	1.77	2.10	1.76	0.04	1.77	1.77	1.77
S.E. (logit-sens)	0.38	0.47	0.07	0.44	0.41	0.48	0.38	0.01	0.38	0.38	0.38
Specificity	0.80	0.83	0.02	0.82	0.81	0.84	0.81	0.01	0.81	0.80	0.82
T ² _{Spec}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
Logit-spec	1.41	1.61	0.17	1.54	1.46	1.64	1.45	0.05	1.46	1.41	1.51
S.E. (logit-spec)	0.22	0.28	0.05	0.26	0.24	0.28	0.23	0.01	0.23	0.23	0.23
70%		Missing data =9.54%					Missing data =0%				
Sensitivity	0.85	0.85	0.00	0.85	0.85	0.85	0.85	0.01	0.85	0.84	0.85
T ² _{Sens}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Logit-sens	1.77	1.77	0.00	1.77	1.76	1.77	1.70	0.07	1.77	1.63	1.77
S.E. (logit-sens)	0.38	0.41	0.03	0.38	0.38	0.44	0.37	0.01	0.38	0.36	0.38
Specificity	0.85	0.86	0.01	0.85	0.85	0.87	0.85	0.00	0.85	0.85	0.85
T ² _{Spec}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Logit-spec	1.74	1.82	0.08	1.74	1.74	1.90	1.73	0.02	1.74	1.74	1.74
S.E. (logit-spec)	0.25	0.27	0.02	0.25	0.25	0.30	0.25	0.00	0.25	0.25	0.25
80%		Missing data 0=0%					Missing data =0%				
Sensitivity	0.73	0.73	0.00	0.73	0.73	0.73	0.73	0.00	0.73	0.73	0.73
T ² _{Sens}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Logit-sens	0.98	0.98	0.00	0.98	0.98	0.98	0.98	0.00	0.98	0.98	0.98
S.E. (logit-sens)	0.30	0.30	0.00	0.30	0.30	0.30	0.30	0.00	0.30	0.30	0.30

Parameter of interest	Mean values from complete data	Meta-analysis with data missing					Meta-analysis with imputed data				
% PTH Decrease		Mean	S.d.	Median	Lower quartile	Upper quartile	Mean	S.d.	Median	Lower quartile	Upper quartile
Specificity	0.91	0.91	0.00	0.91	0.91	0.91	0.91	0.00	0.91	0.91	0.91
T^2_{Spec}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Logit-spec	2.26	2.26	0.00	2.26	2.26	2.26	2.26	0.00	2.26	2.26	2.26
S.E. (logit-spec)	0.30	0.30	0.00	0.30	0.30	0.30	0.30	0.00	0.30	0.30	0.30

*Average missing data across all 1000 datasets

6.5 Discussion

This chapter has found through a simulation study that the imputation method proposed by Riley et al. [158] appears to work well for the PTH example; in general, when missing values are generated, the mean meta-analysis values are brought closer to the true known values after imputation for missing sensitivity and specificity. Also the standard errors of the pooled logit-sensitivity and logit-specificity are closer to the true values after imputation, and there is substantial gain in precision by using additional imputed results. These findings were consistent across the different scenarios for generating missing data as investigated in this chapter. Although more work is required and other different datasets and scenarios should be investigated in the future, the initial findings appear to be promising.

6.5.1 Strengths and limitations of simulation study

Strength of this simulation study is in having the IPD available, as the observed meta-analysis results at each threshold were known in a real dataset; thus the particular relationship between threshold value and sensitivity and specificity did not need to be specified, as the data was already available. In other words, this simulation study uses the real data so there was no need to make assumptions about the underlying ROC shape or assumption of any particular distribution (a downside of this is that the 'true' values were unknown, and so results are compared to the true estimates from complete data). In addition, the method of Riley et al. [158] had not previously been evaluated through a simulation study. This simulation study has also considered a few scenarios to reflect the availability of thresholds in real life, i.e. selective reporting. The scenarios considered ranged from certain probability of missing data, i.e. probability of 0.5 of thresholds being missing in a dataset to missing at random (MAR) to reflect those

instances where thresholds are presented based on how positive the result is (a sensitivity ≥ 0.90).

A key limitation of the findings is that the results are all based on one dataset. This is one case study, and it is not possible to say whether the same results would be achieved in another case study. There was no heterogeneity at most thresholds in the Noordzij data, so further consideration in datasets with larger tau-squared values is required. Furthermore, the true values of sensitivity and specificity (exact values) were not known but rather the comparison was with true estimates, as noted above.

6.5.2 Strengths and limitations of the imputation method

A key strength of the imputation method is the ability to gain additional information for meta-analysis, which allows more studies to be included in a meta-analysis at each threshold. Due to it being a single imputation approach, it is simple and allows standard methods of meta-analysis to be used. This method offers a valuable opportunity for a sensitivity analysis to be performed; to assess whether the conclusions about heterogeneity, best threshold and test accuracy hold. This method appears to give similar estimates on average to the observed estimates, with less variability from the observed estimates than before imputation. Also, in the Apgar data it was flagged that the results may be better than originally thought after using the imputation method. If the standard error is smaller and the bias is the same, the mean squared error (MSE) must be better in the imputation method. Coverage could not be evaluated as the truth was not known.

An important limitation of this method is that it only uses a single imputation approach. As well as standard errors being potentially too small, and an assumption is made that

between two known thresholds the relationship is linear, which is impossible to validate. This method also does not force thresholds to be ordered and no imputation takes place at end thresholds or above/below the upper/lower observed thresholds. Extrapolation may be possible using further assumptions, for thresholds slightly above or below the observed range.

6.5.3 Multivariate meta-analysis model approach

Another approach to deal with multiple thresholds has been discussed by Riley et al. [158]. A multivariate meta-analysis model is used that assumes logit-sensitivity and logit-specificity follow a multivariate normal distribution, both within and between-studies; this increases the information toward each threshold's meta-analysis by utilising within-study and between-study correlation across thresholds. But, this approach may need continuity corrections and can poorly estimate the correlations at +1 or -1. This model is being used increasingly to synthesise multiple outcomes which are correlated, in meta-analysis [163] and is known to improve the efficiency of summary estimates [139]. As well as reducing the outcome reporting bias issues [140].

There are a number of limitations have been discussed for this method. This method involves a time-consuming first step which involves obtaining logit-sensitivity, logit-specificity, their variances and covariance's for each threshold in each study. When a zero cell arises in a 2 by 2 table then a continuity correction is required; also a within-study correlation of 1 can arise which needs to be reduced or the results may not converge in the second step. In the second step, logit-sensitivity and logit-specificity both require an assumption of a within-study multivariate normal sampling distribution, which can result in lower estimation properties when compared to modelling the exact binomial distribution [153]. Another limitation to note is when the between-study

correlation for each pair of thresholds and each pair of sensitivity and specificity are poorly estimated at +1 or -1, this can inflate the between-study variance estimates, although the summary statistics remain unaffected and unbiased [160].

Due to these limitations, the imputation approach is potentially more appealing.

6.5.4 Other methods

Hamza et al. [148] proposed a multivariate random-effect meta-analysis approach and this applies to when all studies report all thresholds. In comparison, the Riley method allows a different set of thresholds per study and can also use studies which only provide one threshold. The Hamza et al. [148] approach models the linear relationship between threshold value and test accuracy within each study, which is not possible in those studies that report only one threshold. This method is susceptible to convergence issues, which prompted Putter et al. [135] to propose an alternative survival model framework for meta-analysing the multiple thresholds. But the model also requires multiple thresholds to be available in all studies. The methods proposed by Hamza et al. [148] or Putter et al. [135] both offer a more sophisticated option when there is a complete set of thresholds in all studies (or when IPD is available for all studies). These methods have been found to have estimation and convergence issues for data that contained missing thresholds across studies [135, 158].

The imputation method proposed by Riley et al. [158] does not assume a common linear relationship between observed pairs of thresholds, and allows different linear relationships in each pair. Therefore the method only assumes a straight line relationship between each pair of observed thresholds available. This is different in comparison to the Hamza et al. [148] approach where they assume a linear relationship

in logit ROC space. A single linear trend is assumed across the whole logit ROC space only when producing the summary ROC curve [163] through a model, to then constrain the summary estimates to be ordered.

6.5.5 Conclusion

Riley et al. [158] proposed a new meta-analysis approach to deal with multiple thresholds per study, and in this chapter this was evaluated through the simulation study. Across the four different scenarios, the general message is that the imputation approach appears suitable at least as a sensitivity analysis to assess how meta-analysis results change after imputation. If thresholds are missing at random, the mean values for pooled logit-sensitivity and logit-specificity are the same or very similar on average compared to the true estimated value. The standard errors are larger in the meta-analysis before imputation, and the meta-analysis after imputation brings these values closer to the true estimated standard errors. The median tau-squared values are very similar.

For selective reporting scenarios, it removes the upwards bias in pooled results and reduces standard errors. In scenario III, although bias in most thresholds was removed on average, for some thresholds the imputation method was slightly too conservative (sensitivity and specificity reduced slightly too far).

This method could potentially help meta-analysts to quickly evaluate the impact of missing and selectively reporting threshold results. This approach uses commonly known methods for diagnostic test accuracy meta-analysis and offers a more practical approach to more sophisticated methods [135, 148] that require complete data or experience convergence issues.

6.5.6 What further research is needed?

Further research is required to validate this imputation method. For example different types of datasets need to be used in simulation case studies such as this. A more heterogeneous dataset may give different results; it would be of interest to find out if the results are better, worse or in fact similar. Many more validation studies are required to then consider recommending this imputation method for wider use. But, in the longer term the method requires extension to multiple imputations.

6.5.7 What this chapter adds?

A summary of what the problem is, what this study adds and what is required next is described in Table 6.9.

Table 6.9: What chapter six adds?

What is known / what is the problem?
<ul style="list-style-type: none">• When performing a meta-analysis of thresholds presented in papers, there will be missing data as different papers will present different thresholds• Rarely do all papers report all thresholds, so it becomes difficult to get a clear view of what the best threshold is or how it compares to the others• An imputation approach has been considered in this chapter, which assumes a linear relationship between two known thresholds and imputes the threshold• The data from Noordzij is used here; as all of the IPD is available, missing thresholds are generated to replicate scenarios in practice
What this study adds?
<ul style="list-style-type: none">• From the imputed results for sensitivity and specificity we can see the bias is reduced in comparison to having missing values, and is much closer to when we have the full data• Standard errors are also reduced and as a result the mean squared error is reduced• Thus gaining more information than compared to having missing values• This imputation approach is a good tool as a sensitivity analysis to assess how good the conclusions are compared to having missing threshold values, and may even change the conclusions depending on the results
What is needed next?
<ul style="list-style-type: none">• More analysis is required, i.e. this dataset had almost no heterogeneity, a more heterogeneous dataset may not work as well in this imputation method and this requires further investigation• Also, a multiple imputation approach should be considered, as this approach used single imputation• There are many articles on publication bias and missing data, this is therefore an exploratory tool to assess whether the conclusions are robust due to the missing data• If imputation approach indicates conclusions are vulnerable it may be necessary to obtain the IPD to then calculate the thresholds with complete data

CHAPTER 7: DISCUSSION

Systematic reviews and meta-analysis of risk prediction research are an important part of clinical research. This thesis has contributed to the growing body of work in this field, through application, development and evaluation of novel statistical methods for meta-analysis. In this final chapter, the key findings of the thesis are summarised and further research needs are addressed.

7.1 Overview of the thesis

Chapter 2 evaluated the benefit of a meta-analysis of aggregate data for summarising and comparing the performance of risk prediction models using a real example in breast cancer. A systematic review was performed and an attempt was made to meta-analyse the reported risk prediction models. Unfortunately, this was not possible due to a lack of complete reporting of the models proposed. It was thus decided to compare the models by performing a meta-analysis of their validation statistics (discrimination and calibration). Clear recommendations were difficult to form, as the models were generally validated in different datasets and validation statistics were often incompletely reported. Thus a key finding is that further external validation studies are required, which are powered and designed to allow all the competing models to be formally compared directly to each other.

Given the difficulties of evaluating prediction models based on published aggregate data, Chapter 3 systematically reviewed the methods and reporting used in existing risk prediction model research that utilised IPD from multiple studies. The systematic review was based on an extensive list of questions, and showed that most articles

disregard clustering within their IPD (from multiple studies) and usually only validate their models internally, although a few studies used an 'internal-external cross-validation' approach to maximise the data toward model development and external validation. The findings led to recommendations for improving the field, and a number of methodological questions were identified. It is important to note that the review only considered articles that develop and/or validate a prediction model. For risk prediction models to become more common in practice, research also needs to show they have a positive impact on health outcomes. Such impact studies are currently rare [19], but they should follow any IPD meta-analysis that develops and validates an accurate risk prediction model.

A key methodology issue was addressed in Chapters 4 and 5: what is the impact of accounting for (versus ignoring) clustering of patients within studies, when using IPD from multiple studies to summarise the performance of a risk prediction tool? For simplicity, the prediction tool considered was a prognostic test (i.e. a model with a single predictor, dichotomised at a particular threshold). A prognostic test dataset was obtained from one of the papers from the systematic review in Chapter 3, which had ignored clustering. In the example, there was almost no heterogeneity in the sensitivity, specificity and C statistic values and the original conclusions generally appeared fine (although at some thresholds the CI's for sensitivity and specificity were smaller by ignoring clustering) based on these values as the unstratified and meta-analysis approaches had similar values. In contrast there appeared to be a problem with calibration of post-test probabilities (PPV and NPV) as there was heterogeneity in the prevalence, and so NPV and PPV based on the summary prevalence, sensitivity and specificity was poor in some studies. However, when summary sensitivity and

specificity were combined with a study's own prevalence, calibration was dramatically improved. Thus, in contrast to the original findings reported, the NPV and PPV for new populations should be tailored to each studies own population [156, 165], and not use the NPV and PPV based solely on summary meta-analysis results. Willis et al. [156] have also had similar findings about the need to check whether meta-analysis results for a particular test are reliable enough to be used in clinical practice. They suggest that tailoring meta-analysis results to particular clinical settings will improve the reliability (validity) of test accuracy results.

Chapter 6 continued to examine meta-analysis for a single prognostic test, and how to deal with published aggregate data where the reported thresholds differ across studies. Simulation and IPD from real studies was used to evaluate a recently proposed method by Riley et al. for imputing missing threshold results in a study, which are bounded between two known thresholds. In general, it was found this method performs well as an exploratory (sensitivity) analysis, as the imputation approach appears to give very little bias in pooled estimates at each threshold, but with substantially increased precision compared to the original data with missing values. Further research is needed to extend the method to multiple imputations, and thereby produce more robust confidence intervals.

7.2 Key findings and recommendations

The work described above contributes important findings to the field of risk prediction research. Four papers have been published (based on the work in Chapter 2 [41], Chapter 3 [16], Chapters 4-5 [43] and Chapter 6 [43]), all the cover pages have been provided in APPENDIX D. The key findings and recommendations have been outlined

at the end of each chapter, and the major implications are summarised once more in Table 7.1.

Table 7.1: Key findings and recommendations

Key findings
<ul style="list-style-type: none"> • Breast cancer risk prediction models are poorly reported, in terms of both model parameter estimates and validation results, and thus it is difficult to perform meta-analysis to summarise current evidence • IPD is needed in meta-analysis of risk prediction research, so that model estimates and validation statistics can be calculated and compared easily • In general, researchers currently ignore clustering of patients within studies when IPD is used to develop and/or validate risk prediction models from multiple studies • Researchers currently validate models mainly using internal validation, but an approach called internal-external cross-validation appears promising to overcome this • A real example of a prognostic test meta-analysis shows that it is very important to take clustering into account when <ul style="list-style-type: none"> • (i) assessing heterogeneity across studies for discrimination (sensitivity, specificity and C statistic), in order to recommend an appropriate threshold to be used across different settings; and • (ii) assessing heterogeneity in post-test probabilities (PPV and NPV); in particular, if there is heterogeneity in the prevalence then calibration will be poor, and so summary meta-analysis results will need tailoring to particular populations according to their own prevalence • When dealing with missing threshold results in an aggregate data meta-analysis of the accuracy of a prognostic test, a single imputation method works well as an exploratory analysis, as it increases precision of pooled estimates and has very little bias, compared to the original analysis that ignores missing thresholds in the data
Recommendations
<ul style="list-style-type: none"> • Papers reporting risk prediction models need to improve their quality of reporting, especially in terms of reporting the actual model (with intercept term (if applicable) and all variables with a 95% CI) • When risk prediction models are validated, their validation statistics should be presented with their uncertainty • It is important for IPD to be made available so when models are not presented or validation statistics are not presented, then one can calculate these with the IPD and directly compare competing models • It is essential to account for clustering and assess heterogeneity when using IPD from multiple studies to develop and/or validate a risk prediction model • Decisions on the best threshold for a prognostic test, and how to implement it in clinical practice, should evaluate test performance in new populations. In particular, researchers should check whether there is any heterogeneity in test accuracy and post-test probabilities • When dealing with meta-analysis of test accuracy studies with missing thresholds, researchers should use the imputation method of Riley et al. as an exploratory analysis to assess the potential impact of missing threshold data on their meta-analysis conclusions

7.3 Main messages for the risk prediction field

This thesis has established through empirical examples and case studies that evaluating risk prediction models across multiple studies is very difficult without IPD. IPD enables models to be derived and compared, with external validation performance statistics calculated and summarised across studies in a meta-analysis. However, Chapter 4 showed that even when IPD are available, problems may remain and researchers often do not make most use of the data at hand. Potential flaws were identified and ways to improve this recommended, in particular assessing discrimination and calibration in each study separately rather than performing these analyses disregarding clustering. This was shown to be pivotal for the PTH test, as the PPV and NPV proposed by Noordzij et al. were found to be unreliable, as they needed to be tailored to new populations according to their own prevalence. Leeflang et al. [166] have used a bivariate random-effects model and have not mentioned whether correlation was poorly estimated at ± 1 or not. If it was, then it would have been better to use the univariate model. They have jointly analysed PPV and NPV, and state that prevalence will be the driving force behind correlation [166]. Whereas in this thesis I have tailored the PPV and NPV values using each studies observed prevalence, which is going beyond what they have done. Willis et al. [156] use a tailored meta-analysis approach where they only include studies in to the meta-analysis if the study fits within or is close-enough to a applicable region in the ROC space [156], this will exclude those studies that are deemed to be not in the applicable region. This is in contrast to the approach used in this thesis, where summary estimates are tailored to the individual study to improve calibration of the predictive test.

Although Chapters 4 to 6 focused on a single prognostic test, many of the issues raised are also applicable to multivariate prediction models (that contain multiple predictors). In particular, the internal-external cross-validation approach was identified as a key method.

Debray et al. [82] develop a multivariable logistic regression model from an IPD meta-analysis with potential between-study heterogeneity. They propose strategies for choosing a valid model intercept for when the model is to be validated or applied to new individuals or study samples. This is similar in concept to using a study specific prevalence. Their results indicate that stratified estimation of model intercepts facilitates the derivation of a study specific model intercept, even when it is to be applied to a new study that was not considered during model development. Where the intercept can be well-matched to the excluded study, Debray et al. [103] show that their framework allows an IPD meta-analysis to produce a single, integrated prediction model that can be implemented in practice and has improved model performance and generalisability. This echoes other recommendations to account for clustering in an IPD meta-analysis [102]. For survival data, this means that the baseline hazard should be modelled during model development and so researchers should move away from using Cox regression (a common approach used which does not estimate the baseline hazard) and rather use other approaches such as flexible parametric methods, like the Royston-Parmar model [101] that estimates the baseline hazard using restricted cubic splines [11, 18].

Klaveren et al. [27] have assessed the discriminative ability of a risk model in clustered data, they have used different methods to meta-analyse Harrell's c-index and found that a 95% prediction interval is quite wide (0.60, 0.95). They have recommended to meta-analyse the c-index at a cluster level, where it is worked out in each cluster separately

and then meta-analysed. This is similar to what I have done in Chapter 4 and found the results improved when performing a meta-analysis that accounts for clustering.

Pennells et al. [130] have assessed a risk prediction model using IPD from multiple studies, they have assessed its' predictive ability by using discrimination measures such as the C statistic. What they have not done is assessed its ability in making individualised predictions, looking at how well it predicts at a study level and how it calibrates. As found in this thesis, discrimination was found to be good, but calibration was poor until it was tailored to each study using their observed prevalence which is more important.

Cochrane currently advise to use summary values for sensitivity, specificity, PPV and NPV from a meta-analysis. This may need to change if they want to do prognostic meta-analysis in the future [167], this is due to the findings of this research in Chapters 4 and 5, which have shown summary values may discriminate well, but calibration is poor. But in another study with more heterogeneity, discrimination may be poor too, thus having to tailor the sensitivity and specificity values to each individual setting as well as the prevalence.

7.4 Further research

One of the main limitations of this thesis is that only one dataset was used within Chapters 4-6, to examine the impact of ignoring clustering and for evaluating the single imputation method. This dataset had almost zero heterogeneity, and there was only one predictor included. It would be interesting to examine whether findings are similar in more heterogeneous datasets in further research. Indeed, in some way the PTH dataset is probably not typical of what we would encounter in the real world; in general we would

find more heterogeneity within the dataset, and be focused on more than one predictor for inclusion in a multivariable prediction model. Nonetheless, although there are limitations of this dataset, it is a real clinical example where adjustment for clustering did not occur in the original research study.

The dataset also did not suffer importantly from missing data. As missing outcome data can cause a big issue when developing a risk prediction model, further research needs to evaluate the issues considered in this context also. Recently an approach has been published by Burgess et al. [168] where they attempt to perform multiple imputation and meta-analysis together, although this requires further research.

Other work that may be considered subsequent to this thesis are to perform an updated systematic review up until 2015 (update of Chapter 3), to assess whether anything has changed since the previous review of IPD meta-analysis of risk prediction studies up to 2009. Perhaps standards have improved, for example clustering might be accounted for more often, or more authors are using the ‘internal-external’ cross-validation approach.

Also, the imputation method needs to be robustly assessed using a more heterogeneous dataset; this will help gain a better insight as to how the imputation method works in a different setting. As well as using a multiple imputation approach instead of single imputation, as single imputation is a simple procedure that fills in the missing value between two known thresholds in this instance and importantly regards this imputed value as an equal to the data that has not been imputed. Whereas multiple imputations will simulate the possible missing values and the uncertainty around them, variability is taken in to account in the imputed data to find a range of suitable imputable values.

Another extension is to extrapolate beyond the known thresholds. In this dataset if 50%

and 70% were only available in a missing data scenario, then only 60% and 65% would be imputed. But, if we assumed that the relationship between all thresholds was linear, then we could extrapolate out to 40% and 90% and even beyond those values (this could again be assessed as the full IPD is available). Riley et al. [169] have presented a multivariate-normal method [148] which deals with multiple and missing thresholds for studies in a meta-analysis, they use the correlation between thresholds to reduce the impact of missing thresholds and produce an ROC curve, which is in contrast to the method used here which performs a separate meta-analysis at each threshold. This method has previously been recommended by Hamza et al. [148] and the approach of meta-analysing with a multiple number of thresholds has been discussed by Dukic et al. [170] and Putter et al. [171].

Also, Chapters 4 to 6 looked at a threshold specific result. This required the continuous predictor to be dichotomised. An extension would be to re-assess the work performed in those chapters with them left as continuous, with the possibility of modelling non-linear trends. Also note, percentage change from baseline has been used for the analysis in this thesis. But we could include two predictors instead, one of baseline and one of the final score in the model to assess whether this is more powerful in comparison. Percentage change is used to reduce impact of different methods of measuring PTH across studies. More research is needed on how to deal with different methods of measurement.

Table 7.2: Further research recommendations

Recommendations
<ul style="list-style-type: none">• Perform an updated systematic review up until 2015 for Chapter 3 to examine if standards have improved• Use a more heterogeneous dataset to assess impact of discrimination and calibration statistics when ignoring clustering• Assess the robustness on the single imputation method using a more heterogeneous dataset• Develop a multiple imputation approach instead of single imputation for dealing with missing thresholds in test accuracy meta-analysis• Extend the imputation methods by potentially assuming a linear relationship and extrapolating beyond known thresholds• Assess PTH as a prognostic test on its continuous scale instead of dichotomising into a binary predictor• Instead of % change from baseline, develop a multivariable model for risk of hypocalcaemia that uses at least two predictors, for example PTH at baseline and PTH after surgery

7.5 Conclusions

This thesis has shown through several chapters that IPD is crucial in the development and validation of risk prediction models and prognostic tests using multiple studies. IPD allows more robust model development, an internal-external cross-validation approach, the estimation and synthesis of validation performance statistics, and the identification of pertinent thresholds of implementing tests. When IPD are not available, the thesis has also shown the practical issues of a systematic review and, if possible, meta-analysis of published results, and empirically evaluated an imputation method for dealing with missing threshold results. Researchers wishing to undertake risk prediction research using multiple studies are highly recommended to seek IPD, and ensure they consider clustering in their analysis and assess heterogeneity in prediction performance, as this is the best approach to produce the most robust risk prediction findings that can be translated into clinical practice.

APPENDIX A

APPENDIX A.1: List of questions for systematic review

1) Background information

What country is the corresponding author located in? (I.e. what is the central location for the IPD project?)

Is there a reference to a *protocol* for the IPD project, and, if so, were details given as to where it can be found?

How the project was funded, and was *ethics approval* granted for the IPD project? If not, were reasons given as to why ethics approval was not necessary and, if so, what were the reasons?

Number of studies/datasets included?

What were the types of different studies included (e.g. studies, databases etc.)?

Number of authors/researchers?

2) Research objectives

What was the key research aims of the paper in relation to prognosis/risk prediction?

At baseline what was the condition of the patients being assessed?
(E.g. what disease did they have, or what operation had they just had, or were they healthy etc., Note: If diagnostic prediction model; than suspicion of the disease is starting point).

What outcomes or diseases were of interest for prediction?

3) Identifying studies and their IPD

What was the process used to identify relevant studies for the IPD project?
(E.g. literature review or collaborative group).

If a literature review, then what search strategy was used (e.g. search of Medline, Embase) and was a list of keywords given? Also was a flow chart shown, giving the flow of studies into and out of the project?

If collaborative group, how were studies or databases chosen to be included in the collaborative group?
(E.g. friends in the field, existing database, etc.)

What are the types of studies for inclusion? e.g. RCTs, just placebo arms from RCTs, cohort study, hospital databases, insurance company registries etc.

Was the total sample size or total number of studies required for the model development and/or validation justified? E.g. sample size calculation, or did they ask for IPD from all studies available?

Also info on number and types of outcomes, as these are related to power considerations together with the number of candidate predictors?

And most importantly perhaps: number of pre-planned subgroup analyses (as this is the main idea for IPDs in at least therapeutic studies)

4) Asking for and obtaining IPD

How were authors of relevant studies approached for IPD (e.g. e-mail, letter, phone etc.)?

How many studies (or collaborating groups) were ultimately approached for IPD, and what proportion of these studies/groups actually provided IPD?

(If appropriate) what were the reasons given as to why some studies refused to provide IPD?

How many studies ultimately provided IPD?

What types of studies were they (e.g. RCT, cohort, etc.)?

Was the number of patients within each of the IPD studies given? If not, was the number of patients across all studies given?

Was the total number of events given for each predicted outcome, within each of the IPD studies? If not, was the total number of events given across all studies?

Was the number of events per candidate variable (predictor) given in each study? If not, was the number of events per variable given across all studies?

Was the number of candidate predictors in total and per study given?

Was a summary of the sample population given for each study separately (e.g. mean age, proportion male, treatments in use etc.); if not, was it given for the whole IPD combined?

Did they synchronise prediction and outcome definitions and measurements?

Did they have to do that?

What attempts were made?

Did they use proxies if not exactly same measurement method or definition?

Did they delete studies due to absence or completely different predictor or outcome?

Did the inclusion / exclusion criteria for an IPD study include an assessment of study quality?

If yes, what quality criteria were used to decide inclusion or exclusion (or ‘low’ quality and ‘high’ quality)?

Were IPD still sought from low quality studies?

5) Missing data

At individual-level: Were details of any missing individual-level data within the available IPD given for each study, and, if so, what were they? (E.g. for some patients their age was unknown) (Note: Check tables to help with this)

At study-level: Did all studies have all candidate predictors or outcomes of interest, or was there any missing data across studies? (E.g. for some studies, age was not recorded at all)

If either of these were detailed, how was missing data handled in the analysis?

6) Model development: statistical analysis methods

Are any articles referred to for methodology or any statistical methods cited? Especially if it relates to using data from multiple studies. (Note: for my own reference mainly)

Was a statistical analysis plan for model development given or mentioned in the Methods section?

Was the number of patients and events used from each study toward the prediction model development given?

Were the prediction models developed using the IPD from multiple studies? If so, how was the data synthesised:

‘One-step ignoring clustering’: lumping all the data together into one big dataset, and ignoring clustering by study or collaborative group; or

‘A one-step analysis accounting for clustering’, where the data from all studies/collaborative groups are analysed together but with clustering by study/group accounted for (e.g. using a dummy variable for study); or

A two-step approach, where the data are first analysed separately in each study, and then their model estimates are pooled together in second-step.

Developed using a part of the dataset, and then validated using the remaining data

What types of statistical models were used?

In the two-step approach, details are needed here of how individual studies were analysed, and then how the model estimates were pooled using meta-analysis.

In the one-step approach, again details are needed here of the one-step model itself (Cox regression, logistic regression) and the meta-analysis assumptions therein (e.g. fixed or random-effects on the predictor effects)

Was between-study heterogeneity in the predictor/outcomes considered within the prediction model? If so how (e.g. using random-effects in the analysis; assessing heterogeneity using I-squared)

How were continuous predictors in the prediction model analysed, on a continuous scale or categorized?

If categorized

Were reasons given as to why this was done?

How many cut-points were used, and how were they chosen?

If on a continuous scale

Were non-linear trends assessed and, if so, how were they modelled? (E.g. splines, fractional polynomials)

Was the continuous factor analysed on its original scale, or was it on a transformed scale and why?

If multivariable models were fitted (i.e. prediction models that included multiple variables)

What criteria were used to decide inclusion of a predictor in the model? (E.g. statistical criteria, such as $p < 0.1$, or clinical criteria such as a hazard ratio > 2 or inclusion of 'smoking' variable regardless)

What selection procedure was used? (E.g. forward, backward, stepwise)

Was the final model given in full (i.e. with parameter estimates and standard error or CI for each)? If not, what was given?

List all the problems that are stated or evident that limited the statistical analysis (E.g. different method of measurements, different predictors available in each study etc.) and how did the authors attempt to overcome these problems?

7) Model validation

Internal validation (using the same data used to generate the model):

Was this done?

If so, how?

How good was the performance of the model after the validation?

External validation (using different data than that used to generate the model)e.g. data from other centres not used to develop the model, or half of the data from each centre that was not used to develop and left to validate instead, etc.)

Was this done?

If so, how?

How good was the performance of the model after the validation?

What discrimination, calibration, reclassification or other statistics (e.g. goodness of fit or R²) were used?

Were any figures given to show model discrimination / accuracy / calibration?

If so, what?

Were the results presented with their CI's or S.E's?

Was the number of studies used in the validation stated?

Was the number of patients and events used from each study toward the prediction model validation given?

8) Dealing with those studies not willing / able to provide their IPD

(If appropriate) Was there an assessment or discussion of whether the available IPD studies were a biased set of all studies in the field?

(If appropriate) for studies not providing IPD, were details given as to the number of patients and events in these studies

(if appropriate) were there any other details provided on the qualitative or quantitative differences between those studies providing IPD and those studies not able to provide IPD? If so, what were these differences?

9) Conclusions and discussion

What were the main clinical conclusions of the IPD project in the Discussion, in relation to the use or implementation of the model?

What limitations and problems of the IPD project were noted in the Discussion?

Were totally different conclusions made given the results?

APPENDIX B

APPENDIX B.1: The results of calibration for 0-20 minutes for the remaining studies

Table for Warren 2002 study, 0-20 minutes

Approach	No. of hypocalcaemia who tested positive			95% CI		No. of hypocalcaemia who tested negative			95% CI		Total with hypocalcaemia			95% CI	
Threshold	E	O	E/O	Lower	Upper	E	O	E/O	Lower	Upper	E	O	E/O	Lower	Upper
(1)	Using the Unstratified prevalence, Unstratified sensitivity and Unstratified specificity														
>40	2.79	3	0.93	0.30	2.88	0.26	1	0.26	0.04	1.86	3.05	4	0.76	0.29	2.03
>50	3.29	3	1.10	0.35	3.40	0.21	1	0.21	0.03	1.47	3.50	4	0.87	0.33	2.33
>60	2.19	3	0.73	0.24	2.27	0.59	1	0.59	0.08	4.17	2.78	4	0.70	0.26	1.85
>65	2.29	3	0.76	0.25	2.36	0.71	1	0.71	0.10	5.05	3.00	4	0.75	0.28	2.00
>70	2.36	3	0.79	0.25	2.44	0.70	1	0.70	0.10	4.97	3.06	4	0.77	0.29	2.04
>72	2.41	3	0.80	0.26	2.49	0.69	1	0.69	0.10	4.90	3.10	4	0.78	0.29	2.07
>80	2.24	2	1.12	0.28	4.48	1.09	2	0.55	0.14	2.18	3.33	4	0.83	0.31	2.22
>90	0.67	0				2.35	4	0.59	0.22	1.56	3.02	4	0.75	0.28	2.01
(2)	Using the univariate prevalence, univariate sensitivity and univariate specificity														
>40	2.93	3	0.98	0.32	3.03	0.25	1	0.25	0.04	1.79	3.18	4	0.80	0.30	2.12
>50	3.34	3	1.11	0.36	3.45	0.21	1	0.21	0.03	1.50	3.55	4	0.89	0.33	2.37
>60	2.45	3	0.82	0.26	2.53	0.56	1	0.56	0.08	4.01	3.01	4	0.75	0.28	2.00
>65	2.57	3	0.86	0.28	2.66	0.69	1	0.69	0.10	4.87	3.26	4	0.81	0.31	2.17
>70	2.62	3	0.87	0.28	2.71	0.68	1	0.68	0.10	4.83	3.30	4	0.82	0.31	2.20
>72	2.73	3	0.91	0.29	2.82	0.67	1	0.67	0.09	4.73	3.40	4	0.85	0.32	2.26
>80	2.26	2	1.13	0.28	4.52	1.05	2	0.53	0.13	2.10	3.31	4	0.83	0.31	2.21
>90	0.59	0				2.78	4	0.70	0.26	1.85	3.38	4	0.84	0.32	2.25
(3)	Using the lower 95% prediction interval bound prevalence, univariate sensitivity and univariate specificity														
>40	1.39	3	0.46	0.15	1.44	0.10	1	0.10	0.01	0.70	1.49	4	0.37	0.14	0.99
>50	1.65	3	0.55	0.18	1.70	0.08	1	0.08	0.01	0.58	1.73	4	0.43	0.16	1.15
>60	1.33	3	0.44	0.14	1.38	0.22	1	0.22	0.03	1.57	1.55	4	0.39	0.15	1.03
>65	1.43	3	0.48	0.15	1.48	0.27	1	0.27	0.04	1.92	1.70	4	0.43	0.16	1.14
>70	1.47	3	0.49	0.16	1.52	0.27	1	0.27	0.04	1.91	1.74	4	0.43	0.16	1.16
>72	1.57	3	0.52	0.17	1.62	0.26	1	0.26	0.04	1.87	1.83	4	0.46	0.17	1.22
>80	1.32	2	0.66	0.17	2.64	0.42	2	0.21	0.05	0.84	1.74	4	0.44	0.16	1.16
>90	0.36	0				1.19	4	0.30	0.11	0.79	1.55	4	0.39	0.15	1.03
(4)	Using the upper 95% prediction interval bound prevalence, univariate sensitivity and univariate specificity														
>40	5.04	3	1.68	0.54	5.21	0.63	1	0.63	0.09	4.47	5.67	4	1.42	0.53	3.78
>50	5.48	3	1.83	0.59	5.66	0.53	1	0.53	0.07	3.76	6.01	4	1.50	0.56	4.00
>60	3.58	3	1.19	0.39	3.70	1.37	1	1.37	0.19	9.75	4.96	4	1.24	0.46	3.30
>65	3.68	3	1.23	0.40	3.81	1.64	1	1.64	0.23	11.66	5.32	4	1.33	0.50	3.55
>70	3.72	3	1.24	0.40	3.84	1.63	1	1.63	0.23	11.57	5.35	4	1.34	0.50	3.56
>72	3.80	3	1.27	0.41	3.93	1.60	1	1.60	0.23	11.35	5.40	4	1.35	0.51	3.60
>80	3.10	2	1.55	0.39	6.19	2.43	2	1.21	0.30	4.85	5.52	4	1.38	0.52	3.68
>90	0.79	0				5.63	4	1.41	0.53	3.75	6.42	4	1.61	0.60	4.28
(5)	Using the studies observed prevalence, univariate sensitivity and univariate specificity														
>40	2.53	3	0.84	0.27	2.61	0.21	1	0.21	0.03	1.46	2.73	4	0.68	0.26	1.82
>50	2.91	3	0.97	0.31	3.01	0.17	1	0.17	0.02	1.22	3.08	4	0.77	0.29	2.05
>60	2.18	3	0.73	0.23	2.26	0.46	1	0.46	0.06	3.27	2.64	4	0.66	0.25	1.76
>65	2.31	3	0.77	0.25	2.39	0.56	1	0.56	0.08	3.99	2.87	4	0.72	0.27	1.91
>70	2.35	3	0.78	0.25	2.43	0.56	1	0.56	0.08	3.95	2.91	4	0.73	0.27	1.94
>72	2.47	3	0.82	0.27	2.55	0.55	1	0.55	0.08	3.87	3.01	4	0.75	0.28	2.01
>80	2.05	2	1.02	0.26	4.10	0.87	2	0.43	0.11	1.73	2.91	4	0.73	0.27	1.94
>90	0.54	0				2.33	4	0.58	0.22	1.55	2.87	4	0.72	0.27	1.91

Where blank, this is due to observed cases being 0 so an E/O ratio was not possible

Table for Warren 2004 study, 0-20 minutes

Approach	No. of hypocalcaemia who tested positive			95% CI		No. of hypocalcaemia who tested negative			95% CI		Total with hypocalcaemia			95% CI	
Threshold	E	O	E/O	Lower	Upper	E	O	E/O	Lower	Upper	E	O	E/O	Lower	Upper
(1) Using the Unstratified prevalence, Unstratified sensitivity and Unstratified specificity															
>40	4.65	3	1.55	0.50	4.81	0.41	0				5.06	3	1.69	0.54	5.23
>50	4.76	3	1.59	0.51	4.92	0.38	0				5.14	3	1.71	0.55	5.31
>60	2.19	2	1.10	0.27	4.39	1.12	1	1.12	0.16	7.96	3.32	3	1.11	0.36	3.43
>65	2.29	2	1.14	0.29	4.57	1.36	1	1.36	0.19	9.65	3.65	3	1.22	0.39	3.77
>70	2.36	2	1.18	0.30	4.72	1.34	1	1.34	0.19	9.48	3.70	3	1.23	0.40	3.82
>72	2.41	2	1.21	0.30	4.82	1.32	1	1.32	0.19	9.36	3.73	3	1.24	0.40	3.86
>80	2.80	2	1.40	0.35	5.60	1.91	1	1.91	0.27	13.55	4.71	3	1.57	0.51	4.87
>90	1.34	0				3.75	3	1.25	0.40	3.88	5.10	3	1.70	0.55	5.27
(2) Using the univariate prevalence, univariate sensitivity and univariate specificity															
>40	4.89	3	1.63	0.53	5.05	0.40	0				5.28	3	1.76	0.57	5.46
>50	4.83	3	1.61	0.52	4.99	0.39	0				5.22	3	1.74	0.56	5.39
>60	2.45	2	1.22	0.31	4.89	1.08	1	1.08	0.15	7.65	3.52	3	1.17	0.38	3.64
>65	2.57	2	1.29	0.32	5.14	1.31	1	1.31	0.18	9.30	3.88	3	1.29	0.42	4.01
>70	2.62	2	1.31	0.33	5.23	1.30	1	1.30	0.18	9.23	3.92	3	1.31	0.42	4.05
>72	2.73	2	1.37	0.34	5.46	1.27	1	1.27	0.18	9.03	4.00	3	1.33	0.43	4.14
>80	2.83	2	1.41	0.35	5.65	1.84	1	1.84	0.26	13.07	4.67	3	1.56	0.50	4.82
>90	1.19	0				4.45	3	1.48	0.48	4.60	5.64	3	1.88	0.61	5.83
(3) Using the lower 95% prediction interval bound prevalence, univariate sensitivity and univariate specificity															
>40	2.32	3	0.77	0.25	2.40	0.15	0				2.48	3	0.83	0.27	2.56
>50	2.38	3	0.79	0.26	2.46	0.15	0				2.53	3	0.84	0.27	2.61
>60	1.33	2	0.67	0.17	2.66	0.42	1	0.42	0.06	3.00	1.75	3	0.58	0.19	1.81
>65	1.43	2	0.72	0.18	2.87	0.52	1	0.52	0.07	3.67	1.95	3	0.65	0.21	2.02
>70	1.47	2	0.74	0.18	2.94	0.51	1	0.51	0.07	3.64	1.98	3	0.66	0.21	2.05
>72	1.57	2	0.78	0.20	3.13	0.50	1	0.50	0.07	3.56	2.07	3	0.69	0.22	2.14
>80	1.65	2	0.83	0.21	3.30	0.74	1	0.74	0.10	5.25	2.39	3	0.80	0.26	2.47
>90	0.71	0				1.91	3	0.64	0.21	1.97	2.62	3	0.87	0.28	2.71
(4) Using the upper 95% prediction interval bound prevalence, univariate sensitivity and univariate specificity															
>40	8.40	3	2.80	0.90	8.68	0.99	0				9.39	3	3.13	1.01	9.71
>50	7.92	3	2.64	0.85	8.18	0.98	0				8.90	3	2.97	0.96	9.20
>60	3.58	2	1.79	0.45	7.16	2.62	1	2.62	0.37	18.62	6.20	3	2.07	0.67	6.41
>65	3.68	2	1.84	0.46	7.36	3.14	1	3.14	0.44	22.26	6.82	3	2.27	0.73	7.05
>70	3.72	2	1.86	0.46	7.43	3.11	1	3.11	0.44	22.08	6.83	3	2.28	0.73	7.06
>72	3.80	2	1.90	0.48	7.60	3.05	1	3.05	0.43	21.68	6.86	3	2.29	0.74	7.09
>80	3.87	2	1.94	0.48	7.74	4.25	1	4.25	0.60	30.15	8.12	3	2.71	0.87	8.39
>90	1.59	0				9.00	3	3.00	0.97	9.31	10.59	3	3.53	1.14	10.95
(5) Using the studies observed prevalence, univariate sensitivity and univariate specificity															
>40	2.70	3	0.90	0.29	2.79	0.18	0				2.89	3	0.96	0.31	2.98
>50	2.75	3	0.92	0.30	2.84	0.18	0				2.93	3	0.98	0.32	3.03
>60	1.52	2	0.76	0.19	3.03	0.50	1	0.50	0.07	3.58	2.02	3	0.67	0.22	2.09
>65	1.63	2	0.81	0.20	3.25	0.62	1	0.62	0.09	4.38	2.24	3	0.75	0.24	2.32
>70	1.67	2	0.83	0.21	3.33	0.61	1	0.61	0.09	4.34	2.28	3	0.76	0.24	2.35
>72	1.77	2	0.88	0.22	3.54	0.60	1	0.60	0.08	4.25	2.37	3	0.79	0.25	2.45
>80	1.86	2	0.93	0.23	3.71	0.88	1	0.88	0.12	6.25	2.74	3	0.91	0.29	2.83
>90	0.80	0				2.25	3	0.75	0.24	2.33	3.05	3	1.02	0.33	3.15

Where blank, this is due to observed cases being 0 so an E/O ratio was not possible

APPENDIX B.2: The results of calibration for 1-2 hours for the remaining studies

Table for Warren 2002 study, 1-2 hours

Approach	No. of hypocalcaemia who tested positive			95% CI		No. of hypocalcaemia who tested negative			95% CI		Total with hypocalcaemia			95% CI	
Threshold	E	O	E/O	Lower	Upper	E	O	E/O	Lower	Upper	E	O	E/O	Lower	Upper
(1)	Using the Unstratified prevalence, Unstratified sensitivity and Unstratified specificity														
>40	1.40	1	1.40	0.20	9.91	0.13	1	0.13	0.02	0.95	1.53	2	0.76	0.19	3.06
>50	1.64	1	1.64	0.23	11.67	0.11	1	0.11	0.02	0.79	1.76	2	0.88	0.22	3.51
>60	1.70	1	1.70	0.24	12.06	0.10	1	0.10	0.01	0.73	1.80	2	0.90	0.23	3.60
>65	1.71	1	1.71	0.24	12.17	0.15	1	0.15	0.02	1.07	1.87	2	0.93	0.23	3.73
>70	0.68	1	0.68	0.10	4.80	0.18	1	0.18	0.03	1.27	0.86	2	0.43	0.11	1.71
>72	0.70	1	0.70	0.10	4.97	0.28	1	0.28	0.04	1.97	0.98	2	0.49	0.12	1.95
>80	0.76	1	0.76	0.11	5.41	0.54	1	0.54	0.08	3.83	1.30	2	0.65	0.16	2.60
>90	0.00	0				1.07	2	0.53	0.13	2.13	1.07	2	0.53	0.13	2.13
(2)	Using the univariate prevalence, univariate sensitivity and univariate specificity														
>40	1.42	1	1.42	0.20	10.06	0.14	1	0.14	0.02	0.98	1.55	2	0.78	0.19	3.11
>50	1.67	1	1.67	0.24	11.84	0.11	1	0.11	0.02	0.81	1.78	2	0.89	0.22	3.56
>60	1.72	1	1.72	0.24	12.18	0.11	1	0.11	0.01	0.75	1.82	2	0.91	0.23	3.64
>65	1.77	1	1.77	0.25	12.57	0.15	1	0.15	0.02	1.08	1.92	2	0.96	0.24	3.84
>70	0.68	1	0.68	0.10	4.84	0.18	1	0.18	0.03	1.30	0.87	2	0.43	0.11	1.73
>72	0.70	1	0.70	0.10	5.00	0.30	1	0.30	0.04	2.10	1.00	2	0.50	0.13	2.00
>80	0.78	1	0.78	0.11	5.55	0.55	1	0.55	0.08	3.90	1.33	2	0.67	0.17	2.66
>90	0.00	0				1.09	2	0.54	0.14	2.18	1.09	2	0.54	0.14	2.18
(3)	Using the lower 95% prediction interval bound prevalence, univariate sensitivity and univariate specificity														
>40	0.69	1	0.69	0.10	4.89	0.05	1	0.05	0.01	0.37	0.74	2	0.37	0.09	1.48
>50	0.85	1	0.85	0.12	6.06	0.04	1	0.04	0.01	0.31	0.90	2	0.45	0.11	1.79
>60	1.01	1	1.01	0.14	7.16	0.04	1	0.04	0.01	0.28	1.05	2	0.52	0.13	2.10
>65	1.06	1	1.06	0.15	7.52	0.06	1	0.06	0.01	0.42	1.12	2	0.56	0.14	2.24
>70	0.45	1	0.45	0.06	3.18	0.07	1	0.07	0.01	0.50	0.52	2	0.26	0.06	1.04
>72	0.47	1	0.47	0.07	3.37	0.11	1	0.11	0.02	0.81	0.59	2	0.29	0.07	1.18
>80	0.58	1	0.58	0.08	4.08	0.22	1	0.22	0.03	1.53	0.79	2	0.40	0.10	1.58
>90	0.00	0				0.44	2	0.22	0.06	0.89	0.44	2	0.22	0.06	0.89
(4)	Using the upper 95% prediction interval bound prevalence, univariate sensitivity and univariate specificity														
>40	2.36	1	2.36	0.33	16.79	0.35	1	0.35	0.05	2.47	2.71	2	1.36	0.34	5.42
>50	2.61	1	2.61	0.37	18.56	0.29	1	0.29	0.04	2.06	2.90	2	1.45	0.36	5.81
>60	2.34	1	2.34	0.33	16.59	0.27	1	0.27	0.04	1.91	2.61	2	1.30	0.33	5.21
>65	2.37	1	2.37	0.33	16.86	0.39	1	0.39	0.05	2.76	2.76	2	1.38	0.35	5.52
>70	0.85	1	0.85	0.12	6.03	0.47	1	0.47	0.07	3.32	1.32	2	0.66	0.16	2.63
>72	0.86	1	0.86	0.12	6.12	0.74	1	0.74	0.10	5.26	1.60	2	0.80	0.20	3.21
>80	0.90	1	0.90	0.13	6.42	1.32	1	1.32	0.19	9.34	2.22	2	1.11	0.28	4.44
>90	0.00	0				2.44	2	1.22	0.30	4.87	2.44	2	1.22	0.30	4.87
(5)	Using the studies observed prevalence, univariate sensitivity and univariate specificity														
>40	1.23	1	1.23	0.17	8.73	0.11	1	0.11	0.02	0.79	1.34	2	0.67	0.17	2.68
>50	1.47	1	1.47	0.21	10.41	0.09	1	0.09	0.01	0.65	1.56	2	0.78	0.19	3.12
>60	1.56	1	1.56	0.22	11.06	0.08	1	0.08	0.01	0.60	1.64	2	0.82	0.21	3.28
>65	1.61	1	1.61	0.23	11.46	0.12	1	0.12	0.02	0.88	1.74	2	0.87	0.22	3.47
>70	0.63	1	0.63	0.09	4.50	0.15	1	0.15	0.02	1.06	0.78	2	0.39	0.10	1.57
>72	0.66	1	0.66	0.09	4.67	0.24	1	0.24	0.03	1.71	0.90	2	0.45	0.11	1.80
>80	0.74	1	0.74	0.10	5.27	0.45	1	0.45	0.06	3.19	1.19	2	0.60	0.15	2.38
>90	0.00	0				0.90	2	0.45	0.11	1.80	0.90	2	0.45	0.11	1.80

Where blank, this is due to observed cases being 0 so an E/O ratio was not possible

Table for Warren 2004 study, 1-2 hours

Approach	No. of hypocalcaemia who tested positive			95% CI		No. of hypocalcaemia who tested negative			95% CI		Total with hypocalcaemia			95% CI	
Threshold	E	O	E/O	Lower	Upper	E	O	E/O	Lower	Upper	E	O	E/O	Lower	Upper
(1) Using the Unstratified prevalence, Unstratified sensitivity and Unstratified specificity															
>40	4.88	3	1.63	0.53	5.05	0.27	0				5.15	3	1.72	0.55	5.33
>50	4.93	3	1.64	0.53	5.10	0.26	0				5.19	3	1.73	0.56	5.37
>60	5.09	3	1.70	0.55	5.27	0.25	0				5.34	3	1.78	0.57	5.52
>65	5.14	3	1.71	0.55	5.31	0.37	0				5.51	3	1.84	0.59	5.69
>70	4.06	3	1.35	0.44	4.20	0.40	0				4.46	3	1.49	0.48	4.61
>72	4.20	3	1.40	0.45	4.34	0.62	0				4.82	3	1.61	0.52	4.98
>80	3.05	2	1.52	0.38	6.09	1.32	1	1.32	0.19	9.37	4.37	3	1.46	0.47	4.51
>90	0.00	0				2.77	3	0.92	0.30	2.86	2.77	3	0.92	0.30	2.86
(2) Using the univariate prevalence, univariate sensitivity and univariate specificity															
>40	4.96	3	1.65	0.53	5.13	0.27	0				5.23	3	1.74	0.56	5.41
>50	5.01	3	1.67	0.54	5.17	0.26	0				5.27	3	1.76	0.57	5.45
>60	5.15	3	1.72	0.55	5.32	0.26	0				5.40	3	1.80	0.58	5.58
>65	5.31	3	1.77	0.57	5.49	0.37	0				5.68	3	1.89	0.61	5.87
>70	4.09	3	1.36	0.44	4.23	0.41	0				4.50	3	1.50	0.48	4.65
>72	4.23	3	1.41	0.45	4.37	0.66	0				4.88	3	1.63	0.52	5.05
>80	3.13	2	1.56	0.39	6.25	1.34	1	1.34	0.19	9.53	4.47	3	1.49	0.48	4.62
>90	0.00	0				2.83	3	0.94	0.30	2.92	2.83	3	0.94	0.30	2.92
(3) Using the lower 95% prediction interval bound prevalence, univariate sensitivity and univariate specificity															
>40	2.41	3	0.80	0.26	2.49	0.11	0				2.52	3	0.84	0.27	2.60
>50	2.56	3	0.85	0.28	2.65	0.10	0				2.66	3	0.89	0.29	2.75
>60	3.03	3	1.01	0.33	3.13	0.10	0				3.12	3	1.04	0.34	3.23
>65	3.18	3	1.06	0.34	3.29	0.14	0				3.32	3	1.11	0.36	3.43
>70	2.69	3	0.90	0.29	2.78	0.16	0				2.85	3	0.95	0.31	2.94
>72	2.85	3	0.95	0.31	2.94	0.25	0				3.10	3	1.03	0.33	3.20
>80	2.30	2	1.15	0.29	4.60	0.53	1	0.53	0.07	3.75	2.83	3	0.94	0.30	2.92
>90	0.00	0				1.15	3	0.38	0.12	1.19	1.15	3	0.38	0.12	1.19
(4) Using the upper 95% prediction interval bound prevalence, univariate sensitivity and univariate specificity															
>40	8.28	3	2.76	0.89	8.55	0.70	0				8.97	3	2.99	0.96	9.28
>50	7.84	3	2.61	0.84	8.11	0.68	0				8.52	3	2.84	0.92	8.81
>60	7.01	3	2.34	0.75	7.24	0.65	0				7.66	3	2.55	0.82	7.92
>65	7.12	3	2.37	0.77	7.36	0.94	0				8.07	3	2.69	0.87	8.34
>70	5.10	3	1.70	0.55	5.27	1.04	0				6.14	3	2.05	0.66	6.34
>72	5.18	3	1.73	0.56	5.35	1.65	0				6.82	3	2.27	0.73	7.05
>80	3.62	2	1.81	0.45	7.23	3.22	1	3.22	0.45	22.83	6.83	3	2.28	0.73	7.06
>90	0.00	0				6.33	3	2.11	0.68	6.54	6.33	3	2.11	0.68	6.54
(5) Using the studies observed prevalence, univariate sensitivity and univariate specificity															
>40	2.80	3	0.93	0.30	2.89	0.13	0				2.92	3	0.97	0.31	3.02
>50	2.95	3	0.98	0.32	3.05	0.12	0				3.07	3	1.02	0.33	3.17
>60	3.40	3	1.13	0.37	3.52	0.12	0				3.52	3	1.17	0.38	3.64
>65	3.56	3	1.19	0.38	3.68	0.17	0				3.73	3	1.24	0.40	3.86
>70	2.96	3	0.99	0.32	3.06	0.19	0				3.15	3	1.05	0.34	3.26
>72	3.12	3	1.04	0.34	3.22	0.30	0				3.42	3	1.14	0.37	3.54
>80	2.48	2	1.24	0.31	4.95	0.63	1	0.63	0.09	4.48	3.11	3	1.04	0.33	3.21
>90	0.00	0				1.37	3	0.46	0.15	1.41	1.37	3	0.46	0.15	1.41

Where blank, this is due to observed cases being 0 so an E/O ratio was not possible

Table for Lombardi study, 1-2 hours

Approach	No. of hypocalcaemia who tested positive			95% CI		No. of hypocalcaemia who tested negative			95% CI		Total with hypocalcaemia			95% CI	
Threshold	E	O	E/O	Lower	Upper	E	O	E/O	Lower	Upper	E	O	E/O	Lower	Upper
(1) Using the Unstratified prevalence, Unstratified sensitivity and Unstratified specificity															
>40	11.16	15	0.74	0.45	1.23	0.42	1	0.42	0.06	3.01	11.59	16	0.72	0.44	1.18
>50	12.33	15	0.82	0.50	1.36	0.39	1	0.39	0.05	2.76	12.72	16	0.80	0.49	1.30
>60	11.89	15	0.79	0.48	1.31	0.44	1	0.44	0.06	3.11	12.33	16	0.77	0.47	1.26
>65	12.00	15	0.80	0.48	1.33	0.65	1	0.65	0.09	4.60	12.65	16	0.79	0.48	1.29
>70	12.86	15	0.86	0.52	1.42	0.64	1	0.64	0.09	4.52	13.49	16	0.84	0.52	1.38
>72	11.21	13	0.86	0.50	1.48	1.08	3	0.36	0.12	1.11	12.28	16	0.77	0.47	1.25
>80	10.66	12	0.89	0.50	1.56	2.22	4	0.56	0.21	1.48	12.88	16	0.81	0.49	1.31
>90	10.00	10	1.00	0.54	1.86	4.37	6	0.73	0.33	1.62	14.37	16	0.90	0.55	1.47
(2) Using the univariate prevalence, univariate sensitivity and univariate specificity															
>40	11.33	15	0.76	0.46	1.25	0.44	1	0.44	0.06	3.09	11.77	16	0.74	0.45	1.20
>50	12.51	15	0.83	0.50	1.38	0.40	1	0.40	0.06	2.82	12.91	16	0.81	0.49	1.32
>60	12.01	15	0.80	0.48	1.33	0.45	1	0.45	0.06	3.19	12.46	16	0.78	0.48	1.27
>65	12.39	15	0.83	0.50	1.37	0.65	1	0.65	0.09	4.64	13.05	16	0.82	0.50	1.33
>70	12.95	15	0.86	0.52	1.43	0.65	1	0.65	0.09	4.63	13.61	16	0.85	0.52	1.39
>72	11.27	13	0.87	0.50	1.49	1.15	3	0.38	0.12	1.19	12.42	16	0.78	0.48	1.27
>80	10.94	12	0.91	0.52	1.61	2.26	4	0.56	0.21	1.50	13.20	16	0.82	0.51	1.35
>90	10.00	10	1.00	0.54	1.86	4.46	6	0.74	0.33	1.65	14.46	16	0.90	0.55	1.48
(3) Using the lower 95% prediction interval bound prevalence, univariate sensitivity and univariate specificity															
>40	5.51	15	0.37	0.22	0.61	0.17	1	0.17	0.02	1.19	5.68	16	0.35	0.22	0.58
>50	6.40	15	0.43	0.26	0.71	0.15	1	0.15	0.02	1.09	6.56	16	0.41	0.25	0.67
>60	7.06	15	0.47	0.28	0.78	0.17	1	0.17	0.02	1.21	7.23	16	0.45	0.28	0.74
>65	7.42	15	0.49	0.30	0.82	0.25	1	0.25	0.04	1.79	7.67	16	0.48	0.29	0.78
>70	8.52	15	0.57	0.34	0.94	0.25	1	0.25	0.04	1.77	8.77	16	0.55	0.34	0.89
>72	7.59	13	0.58	0.34	1.01	0.44	3	0.15	0.05	0.46	8.03	16	0.50	0.31	0.82
>80	8.06	12	0.67	0.38	1.18	0.89	4	0.22	0.08	0.59	8.94	16	0.56	0.34	0.91
>90	10.00	10	1.00	0.54	1.86	1.82	6	0.30	0.14	0.67	11.82	16	0.74	0.45	1.21
(4) Using the upper 95% prediction interval bound prevalence, univariate sensitivity and univariate specificity															
>40	18.92	15	1.26	0.76	2.09	1.10	1	1.10	0.16	7.84	20.02	16	1.25	0.77	2.04
>50	19.61	15	1.31	0.79	2.17	1.01	1	1.01	0.14	7.20	20.62	16	1.29	0.79	2.10
>60	16.35	15	1.09	0.66	1.81	1.16	1	1.16	0.16	8.20	17.51	16	1.09	0.67	1.79
>65	16.62	15	1.11	0.67	1.84	1.67	1	1.67	0.23	11.82	18.29	16	1.14	0.70	1.87
>70	16.14	15	1.08	0.65	1.79	1.66	1	1.66	0.23	11.81	17.81	16	1.11	0.68	1.82
>72	13.80	13	1.06	0.62	1.83	2.88	3	0.96	0.31	2.98	16.68	16	1.04	0.64	1.70
>80	12.66	12	1.05	0.60	1.86	5.41	4	1.35	0.51	3.60	18.07	16	1.13	0.69	1.84
>90	10.00	10	1.00	0.54	1.86	9.98	6	1.66	0.75	3.70	19.98	16	1.25	0.77	2.04
(5) Using the studies observed prevalence, univariate sensitivity and univariate specificity															
>40	15.05	15	1.00	0.60	1.66	0.69	1	0.69	0.10	4.92	15.74	16	0.98	0.60	1.61
>50	16.10	15	1.07	0.65	1.78	0.63	1	0.63	0.09	4.50	16.73	16	1.05	0.64	1.71
>60	14.36	15	0.96	0.58	1.59	0.72	1	0.72	0.10	5.11	15.08	16	0.94	0.58	1.54
>65	14.69	15	0.98	0.59	1.62	1.04	1	1.04	0.15	7.41	15.74	16	0.98	0.60	1.61
>70	14.75	15	0.98	0.59	1.63	1.04	1	1.04	0.15	7.41	15.79	16	0.99	0.60	1.61
>72	12.70	13	0.98	0.57	1.68	1.83	3	0.61	0.20	1.89	14.53	16	0.91	0.56	1.48
>80	11.94	12	0.99	0.56	1.75	3.52	4	0.88	0.33	2.34	15.45	16	0.97	0.59	1.58
>90	10.00	10	1.00	0.54	1.86	6.76	6	1.13	0.51	2.51	16.76	16	1.05	0.64	1.71

APPENDIX B.3: The results of calibration for 6 hours

Table for Lam study, 6 hours

Approach	No. of hypocalcaemia who tested positive			95% CI		No. of hypocalcaemia who tested negative			95% CI		Total with hypocalcaemia			95% CI	
Threshold	E	O	E/O	Lower	Upper	E	O	E/O	Lower	Upper	E	O	E/O	Lower	Upper
(1)	Using the Unstratified prevalence, Unstratified sensitivity and Unstratified specificity														
>40	7.43	12	0.62	0.35	1.09	0.25	0				7.68	12	0.64	0.36	1.13
>50	8.21	12	0.68	0.39	1.20	0.24	0				8.45	12	0.70	0.40	1.24
>60	9.08	12	0.76	0.43	1.33	0.23	0				9.31	12	0.78	0.44	1.37
>65	10.46	12	0.87	0.49	1.53	0.22	0				10.68	12	0.89	0.51	1.57
>70	10.92	12	0.91	0.52	1.60	0.42	0				11.35	12	0.95	0.54	1.67
>72	10.66	12	0.89	0.50	1.56	0.65	0				11.31	12	0.94	0.54	1.66
>80	9.60	11	0.87	0.48	1.58	1.32	1	1.32	0.19	9.34	10.92	12	0.91	0.52	1.60
>90	7.22	7	1.03	0.49	2.16	2.62	5	0.52	0.22	1.26	9.84	12	0.82	0.47	1.44
(2)	Using the univariate prevalence, univariate sensitivity and univariate specificity														
>40	7.53	12	0.63	0.36	1.10	0.26	0				7.79	12	0.65	0.37	1.14
>50	8.30	12	0.69	0.39	1.22	0.24	0				8.55	12	0.71	0.40	1.25
>60	9.15	12	0.76	0.43	1.34	0.24	0				9.39	12	0.78	0.44	1.38
>65	10.52	12	0.88	0.50	1.54	0.22	0				10.74	12	0.90	0.51	1.58
>70	10.98	12	0.91	0.52	1.61	0.43	0				11.41	12	0.95	0.54	1.67
>72	10.75	12	0.90	0.51	1.58	0.58	0				11.33	12	0.94	0.54	1.66
>80	9.66	11	0.88	0.49	1.59	1.33	1	1.33	0.19	9.42	10.99	12	0.92	0.52	1.61
>90	7.24	7	1.03	0.49	2.17	2.67	5	0.53	0.22	1.28	9.91	12	0.83	0.47	1.45
(3)	Using the lower 95% prediction interval bound prevalence, univariate sensitivity and univariate specificity														
>40	3.86	12	0.32	0.18	0.57	0.10	0				3.96	12	0.33	0.19	0.58
>50	4.64	12	0.39	0.22	0.68	0.09	0				4.74	12	0.39	0.22	0.70
>60	5.84	12	0.49	0.28	0.86	0.09	0				5.93	12	0.49	0.28	0.87
>65	7.47	12	0.62	0.35	1.10	0.09	0				7.56	12	0.63	0.36	1.11
>70	8.11	12	0.68	0.38	1.19	0.17	0				8.28	12	0.69	0.39	1.21
>72	8.38	12	0.70	0.40	1.23	0.22	0				8.60	12	0.72	0.41	1.26
>80	7.32	11	0.67	0.37	1.20	0.52	1	0.52	0.07	3.71	7.84	12	0.65	0.37	1.15
>90	6.27	7	0.90	0.43	1.88	1.08	5	0.22	0.09	0.52	7.35	12	0.61	0.35	1.08
(4)	Using the upper 95% prediction interval bound prevalence, univariate sensitivity and univariate specificity														
>40	11.78	12	0.98	0.56	1.73	0.67	0				12.45	12	1.04	0.59	1.83
>50	11.84	12	0.99	0.56	1.74	0.63	0				12.47	12	1.04	0.59	1.83
>60	11.66	12	0.97	0.55	1.71	0.61	0				12.27	12	1.02	0.58	1.80
>65	12.44	12	1.04	0.59	1.83	0.58	0				13.02	12	1.08	0.62	1.91
>70	12.68	12	1.06	0.60	1.86	1.11	0				13.78	12	1.15	0.65	2.02
>72	12.04	12	1.00	0.57	1.77	1.46	0				13.51	12	1.13	0.64	1.98
>80	10.99	11	1.00	0.55	1.80	3.20	1	3.20	0.45	22.70	14.19	12	1.18	0.67	2.08
>90	7.69	7	1.10	0.52	2.31	6.06	5	1.21	0.50	2.91	13.76	12	1.15	0.65	2.02
(5)	Using the studies observed prevalence, univariate sensitivity and univariate specificity														
>40	9.68	12	0.81	0.46	1.42	0.42	0				10.10	12	0.84	0.48	1.48
>50	10.17	12	0.85	0.48	1.49	0.39	0				10.56	12	0.88	0.50	1.55
>60	10.55	12	0.88	0.50	1.55	0.38	0				10.93	12	0.91	0.52	1.60
>65	11.62	12	0.97	0.55	1.71	0.36	0				11.98	12	1.00	0.57	1.76
>70	11.96	12	1.00	0.57	1.76	0.69	0				12.66	12	1.05	0.60	1.86
>72	11.51	12	0.96	0.54	1.69	0.92	0				12.43	12	1.04	0.59	1.82
>80	10.44	11	0.95	0.53	1.71	2.07	1	2.07	0.29	14.71	12.51	12	1.04	0.59	1.84
>90	7.51	7	1.07	0.51	2.25	4.07	5	0.81	0.34	1.96	11.59	12	0.97	0.55	1.70

Where blank, this is due to observed cases being 0 so an E/O ratio was not possible

Table for Lombardi study, 6 hours

Approach	No. of hypocalcaemia who tested positive			95% CI		No. of hypocalcaemia who tested negative			95% CI		Total with hypocalcaemia			95% CI	
Threshold	E	O	E/O	Lower	Upper	E	O	E/O	Lower	Upper	E	O	E/O	Lower	Upper
(1) Using the Unstratified prevalence, Unstratified sensitivity and Unstratified specificity															
>40	12.38	15	0.83	0.50	1.37	0.30	1	0.30	0.04	2.12	12.68	16	0.79	0.49	1.29
>50	12.83	15	0.86	0.52	1.42	0.31	1	0.31	0.04	2.22	13.14	16	0.82	0.50	1.34
>60	13.61	15	0.91	0.55	1.51	0.32	1	0.32	0.04	2.26	13.93	16	0.87	0.53	1.42
>65	13.44	15	0.90	0.54	1.49	0.33	1	0.33	0.05	2.36	13.78	16	0.86	0.53	1.41
>70	12.48	14	0.89	0.53	1.51	0.69	2	0.34	0.09	1.37	13.17	16	0.82	0.50	1.34
>72	12.30	13	0.95	0.55	1.63	1.03	3	0.34	0.11	1.07	13.33	16	0.83	0.51	1.36
>80	10.41	11	0.95	0.52	1.71	2.12	5	0.42	0.18	1.02	12.52	16	0.78	0.48	1.28
>90	9.03	10	0.90	0.49	1.68	3.88	6	0.65	0.29	1.44	12.91	16	0.81	0.49	1.32
(2) Using the univariate prevalence, univariate sensitivity and univariate specificity															
>40	12.55	15	0.84	0.50	1.39	0.31	1	0.31	0.04	2.17	12.86	16	0.80	0.49	1.31
>50	12.97	15	0.86	0.52	1.43	0.32	1	0.32	0.05	2.27	13.29	16	0.83	0.51	1.36
>60	13.73	15	0.92	0.55	1.52	0.33	1	0.33	0.05	2.33	14.06	16	0.88	0.54	1.43
>65	13.52	15	0.90	0.54	1.50	0.34	1	0.34	0.05	2.41	13.86	16	0.87	0.53	1.41
>70	12.55	14	0.90	0.53	1.51	0.70	2	0.35	0.09	1.41	13.25	16	0.83	0.51	1.35
>72	12.41	13	0.95	0.55	1.64	0.92	3	0.31	0.10	0.95	13.32	16	0.83	0.51	1.36
>80	10.47	11	0.95	0.53	1.72	2.13	5	0.43	0.18	1.03	12.60	16	0.79	0.48	1.29
>90	9.05	10	0.91	0.49	1.68	3.96	6	0.66	0.30	1.47	13.01	16	0.81	0.50	1.33
(3) Using the lower 95% prediction interval bound prevalence, univariate sensitivity and univariate specificity															
>40	6.43	15	0.43	0.26	0.71	0.12	1	0.12	0.02	0.84	6.55	16	0.41	0.25	0.67
>50	7.26	15	0.48	0.29	0.80	0.12	1	0.12	0.02	0.87	7.38	16	0.46	0.28	0.75
>60	8.76	15	0.58	0.35	0.97	0.12	1	0.12	0.02	0.89	8.89	16	0.56	0.34	0.91
>65	9.61	15	0.64	0.39	1.06	0.13	1	0.13	0.02	0.93	9.74	16	0.61	0.37	0.99
>70	9.27	14	0.66	0.39	1.12	0.27	2	0.13	0.03	0.54	9.54	16	0.60	0.37	0.97
>72	9.67	13	0.74	0.43	1.28	0.35	3	0.12	0.04	0.37	10.02	16	0.63	0.38	1.02
>80	7.93	11	0.72	0.40	1.30	0.84	5	0.17	0.07	0.40	8.77	16	0.55	0.34	0.89
>90	7.83	10	0.78	0.42	1.46	1.60	6	0.27	0.12	0.59	9.43	16	0.59	0.36	0.96
(4) Using the upper 95% prediction interval bound prevalence, univariate sensitivity and univariate specificity															
>40	19.64	15	1.31	0.79	2.17	0.79	1	0.79	0.11	5.58	20.43	16	1.28	0.78	2.08
>50	18.50	15	1.23	0.74	2.05	0.83	1	0.83	0.12	5.87	19.33	16	1.21	0.74	1.97
>60	17.49	15	1.17	0.70	1.93	0.85	1	0.85	0.12	6.01	18.33	16	1.15	0.70	1.87
>65	15.99	15	1.07	0.64	1.77	0.88	1	0.88	0.12	6.27	16.88	16	1.05	0.65	1.72
>70	14.49	14	1.03	0.61	1.75	1.79	2	0.90	0.22	3.58	16.28	16	1.02	0.62	1.66
>72	13.90	13	1.07	0.62	1.84	2.32	3	0.77	0.25	2.40	16.22	16	1.01	0.62	1.65
>80	11.91	11	1.08	0.60	1.95	5.14	5	1.03	0.43	2.47	17.05	16	1.07	0.65	1.74
>90	9.62	10	0.96	0.52	1.79	8.98	6	1.50	0.67	3.33	18.60	16	1.16	0.71	1.90
(5) Using the studies observed prevalence, univariate sensitivity and univariate specificity															
>40	16.13	15	1.08	0.65	1.78	0.49	1	0.49	0.07	3.48	16.62	16	1.04	0.64	1.70
>50	15.89	15	1.06	0.64	1.76	0.51	1	0.51	0.07	3.66	16.41	16	1.03	0.63	1.67
>60	15.82	15	1.05	0.64	1.75	0.52	1	0.52	0.07	3.73	16.35	16	1.02	0.63	1.67
>65	14.94	15	1.00	0.60	1.65	0.55	1	0.55	0.08	3.88	15.49	16	0.97	0.59	1.58
>70	13.67	14	0.98	0.58	1.65	1.12	2	0.56	0.14	2.24	14.80	16	0.92	0.57	1.51
>72	13.28	13	1.02	0.59	1.76	1.46	3	0.49	0.16	1.51	14.75	16	0.92	0.56	1.50
>80	11.31	11	1.03	0.57	1.86	3.33	5	0.67	0.28	1.60	14.64	16	0.92	0.56	1.49
>90	9.39	10	0.94	0.51	1.75	6.04	6	1.01	0.45	2.24	15.43	16	0.96	0.59	1.57

Table for E/O ratios for all the patients combined using the unstratified and univariate meta-analysis approaches at the 6 hour's time-point

% PTH Decrease	No. of hypocalcaemia who tested positive			95% CI		No. of hypocalcaemia who tested negative			95% CI		Total with hypocalcaemia			95% CI	
	E	O	E/O	Lower	Upper	E	O	E/O	Lower	Upper	E	O	E/O	Lower	Upper
<i>Approach 1</i>															
>40	19.81	44	0.45	0.34	0.60	0.61	3	0.20	0.07	0.63	20.42	47	0.43	0.33	0.58
>50	21.04	44	0.48	0.36	0.64	0.61	3	0.20	0.07	0.63	21.64	47	0.46	0.35	0.61
>60	22.69	40	0.57	0.42	0.77	0.61	7	0.09	0.04	0.18	23.30	47	0.50	0.37	0.66
>65	23.90	38	0.63	0.46	0.86	0.61	9	0.07	0.04	0.13	24.51	47	0.52	0.39	0.69
>70	23.41	38	0.62	0.45	0.85	1.23	9	0.14	0.07	0.26	24.64	47	0.52	0.39	0.70
>72	22.96	38	0.60	0.44	0.83	1.85	9	0.21	0.11	0.39	24.81	47	0.53	0.40	0.70
>80	20.01	32	0.63	0.44	0.88	3.77	15	0.25	0.15	0.42	23.78	47	0.51	0.38	0.67
>90	16.25	16	1.02	0.62	1.66	7.10	31	0.23	0.16	0.33	23.36	47	0.50	0.37	0.66
<i>Approach 2</i>															
>40	20.08	44	0.46	0.34	0.61	0.55	3	0.18	0.06	0.57	20.63	47	0.44	0.33	0.58
>50	21.27	44	0.48	0.36	0.65	0.55	3	0.18	0.06	0.57	21.82	47	0.46	0.35	0.62
>60	22.88	40	0.57	0.42	0.78	0.55	7	0.08	0.04	0.16	23.43	47	0.50	0.37	0.66
>65	24.04	38	0.63	0.46	0.87	0.55	9	0.06	0.03	0.12	24.59	47	0.52	0.39	0.70
>70	23.53	38	0.62	0.45	0.85	1.11	9	0.12	0.06	0.24	24.64	47	0.52	0.39	0.70
>72	23.16	38	0.61	0.44	0.84	1.46	9	0.16	0.08	0.31	24.62	47	0.52	0.39	0.70
>80	20.13	32	0.63	0.44	0.89	3.38	15	0.23	0.14	0.37	23.51	47	0.50	0.38	0.67
>90	16.29	16	1.02	0.62	1.66	6.49	31	0.21	0.15	0.30	22.78	47	0.48	0.36	0.65

APPENDIX C

APPENDIX C.1: Stata code used in Chapter 6

**This part of the code is to generate 1000 different datasets each with a different combination of 50% missing data for each threshold, this part is used in all of the scenarios but the code is changed for each scenario depending on what the missing data to be generated is.

**Run through reps, storing estimates

```
local reps=1000 //set reps
```

```
local k=1
```

```
while `k' <= `reps' { //loop through reps to run simulations
```

```
    **generate some data, but first input original data from Noordzij
```

```
        use "C:\Users\Ikhlaaq\Documents\PhD\Chapter 6\1-2hrstudysim.dta", clear
```

```
        gen iteration=`k'
```

```
        **sort by study but keeping thresholds ascending
```

```
        sort studyname, stable
```

```
        by study: generate thresh=_n
```

```
        **gen i and j values for all 35 observations and then for each threshold
```

```
        gen num=_n
```

```
        sort threshold, stable
```

```
        by threshold: generate studyid=_n
```

```
        **sort by study but keeping thresholds ascending
```

```
        sort studyname, stable
```

```
        di " ITERATION `k' "
```

```

**probability of missing at random being 50%

gen x1 = runiform()

gen p=x1

replace x1 = . if p < .5

count if x1!=.

replace tp = . if x1 == .

replace fn = . if x1 == .

replace fp = . if x1 == .

replace tn = . if x1 == .

**This next line calculates how many studies are providing data for each threshold, important for M-A

egen nm = count(tp), by(thresh)

**advance to next replication - starts loop again and stores estimates this time in the next row of your
tempfile

save iteration`k', replace

local k = `k'+1

}

**Merge all the datasets together

use iteration1, replace

for values k=2/1000 {

    append using iteration`k'

}

save originalmissing, replace

**The data has been saved and now I prepare the data for meta-analysis, and also the next section adds 0.5 to each cell to avoid
having any zero cells

use originalmissing, replace

//prep data for analysis

replace tp=(tp+0.5)

replace fn=(fn+0.5)

replace fp=(fp+0.5)

replace tn=(tn+0.5)

save missingdata, replace

**Meta-analysis for missing data, putting the data in to meta-analysis format, loop for 1000 datasets

cd "C:\Users\Ikhlaaq\Documents\PhD\Chapter 6\50%"

use "C:\Users\Ikhlaaq\Documents\PhD\Chapter 6\50%\missingdata.dta", clear

**Convert to meta-analysis format

capture drop n true n1 n0 true1 true0 study sens spec

```

```

gen long n1=tp+fn
gen long n0=fp+tn
gen long true1=tp
gen long true0=tn
gen long study=_n
reshape long n true, i(study) j(sens)

sort study sens

gen byte spec=1-sens

tab nm

save data1, replace

**Meta-analysis code

tempname ests
tempfile estimates_file

postfile `ests' dataset ite sens seSens speci seSpeci lntausens lntauspec using `estimates_file', replace

forvalues h = 1/1000 {
    use data1, replace
    drop if iteration!="'h'
    {
forvalues i = 1/7 {
    save data, replace
    drop if thresh!="'i'
    drop if tp==.
    if nm<2 { *where nm is the number of studies providing data for each threshold
        if nm == 1 {
            gen sensitivity=(tp)/(tp+fn)
            local lnsensitivity=ln(sensitivity/(1-sensitivity))
            local lnsense= 1 / ((tp+fn)*(sensitivity)*(1-sensitivity))
            gen specificity=(tn)/(tn+fp)
            local lnspecificity=ln(specificity/(1-specificity))
            local lnspecse= 1/ ((tn+fp)*(specificity)*(1-specificity))

            local sens = `lnsensitivity'
            local seSens = `lnsense'
            local speci = `lnspecificity'
            local seSpeci = `lnspecse'

```



```

        local Intausens = 0

        local Intauspec = 0

        use data, replace

        }

        else if nm == 0 {

            local sens = .

            local seSens = .

            local speci = .

            local seSpeci = .

            local Intausens = .

            local Intauspec = .

            use data, replace

            }

        }

    else {

        use data, replace

        *Univariate meta-analysis code

        capture xtmelogit true sens spec, nocons || study: sens spec if thresh == `i' & nm>1, nocons binomial(n)

intpoints(1) variance

        local sens = _b[sens]

        local seSens = _se[sens]

        local speci = _b[spec]

        local seSpeci = _se[spec]

        local Intausens = _b[lns1_1_1:_cons]

        local Intauspec = _b[lns1_1_2:_cons]

        }

    local dataset = `h'

    local ite = `i'

    post `ests' (`dataset') (`ite') (`sens') (`seSens') (`speci') (`seSpeci') (`Intausens') (`Intauspec')

    }

    }

    }

postclose `ests'

use `estimates_file', replace

save MD-MA, replace

**Postfile closed, and data saved

```

```

use missingdata,replace

//code for adding the weight to each cut-off

gen weight=0

replace weight=0 if threshold==40

replace weight=2 if threshold==50

replace weight=4 if threshold==60

replace weight=5 if threshold==65

replace weight=6 if threshold==70

replace weight=8 if threshold==80

replace weight=10 if threshold==90

**Missing data indicator

gen missingind=0

replace missingind=1 if tp==.

**generating sensitivity and logit sensitivity for imputation

gen sen=(tp)/(tp+fn)

gen sens=ln(sen/(1-sen))

gen spe=(tn)/(tn+fp)

gen speci=ln(spe/(1-spe))

save missingdata1, replace

**imputation for missing data

capture program drop sim_impute

program define sim_impute, rclass

/* Syntax

            D = Number of datasets

            S = Number of studies

            V = Vector of number of studies per dataset

            T = Number of thresholds

*/

    syntax [, D(int0) S(int 0) V(string) T(int 0)]

/* Create a dataID to identify when a dataset begins and ends*/

gen idnum=_n

if `s'>0 {

    local obs = `s'*`t'

    egen dataID = seq(), b(`obs')

}

```

```

else {

    gen dataID = 1 if idnum<=((`v'[1,1])*`t')

    if `d`>1 {

        local counter = (`v'[1,1])*`t'

        forvalues j = 2/^d' {

            local counter = `counter' + ((`v'[1,`j'])*`t')

            replace dataID = `j' if dataID==. &idnum<=`counter'

        }

    }

}

/* Save whole dataset and drop any extra meta-analytic datasets to enable analysis of each dataset
    individually */

forvalues dataset=1/^d' {

    save data_test, replace

    drop if dataID!=`dataset'

    preserve

/* Check whether user has defined local "s", this indicates that there are the same number of studies
    in each meta-analytic dataset.

    If "s" is not defined then there are a variable number of studies in each meta-analytic dataset
    and the number of studies in each dataset is defined in the matrix name "v" */

    if `s`>0 {

        local ds=`s'

    }

    else {

        local ds=`v'[1,`dataset']

    }

/* Loop through the studies, dropping out all but one study to allow analysis of each study individually*/

forvalues studies=1/^ds' {

    drop if studyid!=`studies'

    local tminus=`t'-1

/* Cycle through the thresholds looking for those without data, which could be imputed*/

    Forvalues i=2/^tminus' {

        if missingind[`i']==1 {

            local down=`i'+1

```

```
/* Cycle down the list of thresholds, looking for the next threshold with data and store this data*/
```

```
forvalues j='down'/'t' {
    if missingind[`j']==0 {
        local DthreshNM= `j'
        local DsensNM= sens[`j']
        local DspeciNM= speci[`j']
        local DweightNM= weight[`j']
        continue, break
    }
    if `j'=='t' {
        local DsensNM= .
        local DweightNM= .
    }
}
```

```
/* Cycle up the list of thresholds, looking for the next threshold with data and store this data*/
```

```
Save looking_for_higher_thresh, replace
drop if thresh>`i'

forvalues k= 1/'i' {
    local real_thresh=(`i'-'k')
    if missingind[`real_thresh']==0 {
        local UthreshNM= `real_thresh'
        local UsensNM = sens[`real_thresh']
        local UspeciNM = speci[`real_thresh']
        local UweightNM = weight[`real_thresh']
        continue, break
    }
    if `real_thresh'==1 {
        local UsensNM= .
        local UweightNM= .
    }
}
```

```
/* Take the weight of the threshold without data, and the data stored from above and below thresholds
```

```
and use this to impute the new threshold data*/

use looking_for_higher_thresh, replace

local w = weight[`i']
```

```

        replace sens=(`UsensNM'+((`DsensNM'-`UsensNM')*(`w'-`UweightNM')/(`DweightNM'-`UweightNM'))))
///
        if thresh=='i'
            replace speci=(`UspeciNM'+((`DspeciNM'-`UspeciNM')*(`w'-`UweightNM')/(`DweightNM'-
`UweightNM'))))///
            if thresh=='i'
                }
            }
/* Save the imputed data file for this study and then restore all studies data from memory*/
    Save imputed_study_data`studies', replace
    restore, preserve
}

/* Append all the study data files for one dataset, then restore the whole dataset and loop round to
analyse the next meta-analytic dataset*/
use imputed_study_data1, replace
if `ds'>1 {
    forvalues a=2/`ds' {
        append using imputed_study_data`a'
        save imputed_dataset_data`dataset', replace
    }
}
restore, not
use data_test, replace
}

/* Append the multiple meta-analytic datasets with imputed data*/
if `ds'==1 {
    use imputed_study_data1, replace
}
else {
    use imputed_dataset_data1, replace

    if `d'>1 {
        forvalues a=2/`d' {
            append using imputed_dataset_data`a'
            save imputed_whole_IPD, replace
        }
    }
}

```

```

    }
}

end

*****

cd "C:\Users\Ikhlaaq\Documents\PhD\Chapter 6\50%"

set more off

use missingdata1, clear

sim_impute, d(1000) s(5) t(7)

save imputeddata, replace

**Imputation has occurred and data has been saved

**put back in to 2*2 format (back-transforming)

use "C:\Users\Ikhlaaq\Documents\PhD\Chapter 6\50%\imputeddata.dta", clear

**indicator for imputed data

gen impind=0

replace impind=1 if missingind==1 &sens!=.

**putting values back in from logitsens and logit spec

replace sen=(exp(sens)/(1+exp(sens))) if impind==1

replace spe=(exp(spec)/(1+exp(spec))) if impind==1

**putting tp, tn, fn and fp values back in

gen hypo=0

gen normo=0

replace hypo=14 if studyid==1

replace normo=28 if studyid==1

replace hypo=18 if studyid==2

replace normo=37 if studyid==2

replace hypo=14 if studyid==3

replace normo=27 if studyid==3

replace hypo=4 if studyid==4

replace normo=10 if studyid==4

replace hypo=5 if studyid==5

replace normo=25 if studyid==5

*These are true disease values

**putting in values for tpetc with imputed sens and spec

replace tp=sen*hypo if impind==1

replace tn=spe*normo if impind==1

```

```

replace fn=hypo-tp if impind==1
replace fp=normo-tn if impind==1

drop sens speci

drop sen spe

save imputeddata1, replace

**calculate nm for the imputed data as this has changed after the imputation has occurred

use imputeddata1, replace

drop nm

save imputeddata1, replace

**generate new nm values for imputed data

**Run through reps, storing estimates

local reps=1000 //set reps

local k=1

while `k' <= `reps' { //loop though reps to run simulations

    set more off

    use "C:\Users\Ikhlaaq\Documents\PhD\Chapter 6\50%\imputeddata1.dta", clear

    drop if iteration!=`k'

    egen nm = count(tp), by(thresh)

    /* advance to next replication - starts loop again and stores estimates this time

        in the next row of your tempfile */

    save iterationind`k', replace

    local k = `k'+1

}

//Merge all the datasets together

use iterationind1, replace

forvalues k=2/1000 {

    append using iterationind`k'

}

save imputeddata2, replace

*data has been updated

//imputed data MA

//loop for 1000 datasets

cd "C:\Users\Ikhlaaq\Documents\PhD\Chapter 6\50%"

*Meta-analysis of Imputed data

use "C:\Users\Ikhlaaq\Documents\PhD\Chapter 6\50%\imputeddata2.dta", clear

```

```

//meta-analysis format

capture drop n true n1 n0 true1 true0 study sens spec

gen long n1=tp+fn

gen long n0=fp+tn

gen long true1=tp

gen long true0=tn

gen long study=_n

reshape long n true, i(study) j(sens)

sort study sens

gen byte spec=1-sens

tab nm

save data1, replace

//set trace on

set more off

tempname ests

tempfile estimates_file

postfile `ests' dataset ite sens seSens speci seSpeci lntausens lntauspec using `estimates_file', replace

forvalues h = 1/1000 {

    use data1, replace

    drop if iteration!=`h'

    {

forvalues i = 1/7 {

    save data, replace

    drop if thresh!=`i'

    drop if tp==.

    if nm<2 {

        if nm == 1 {

            gen sensitivity=(tp)/(tp+fn)

            local lnsensitivity=ln(sensitivity/(1-sensitivity))

            local lnsense= 1 / ((tp+fn)*(sensitivity)*(1-sensitivity))

            gen specificity=(tn)/(tn+fp)

            local lnspecificity=ln(specificity/(1-specificity))

            local lnspecse= 1/ ((tn+fp)*(specificity)*(1-specificity))

            local sens = `lnsensitivity'

```



```

        local seSens = `Insense'

        local speci = `Inspecificity'

        loca lseSpeci = `Inspece'

        local lntausens = 0

        local lntauspec = 0

        use data, replace

    }

    else if nm == 0 {

        local sens = .

        local seSens = .

        local speci = .

        local seSpeci = .

        local lntausens = .

        local lntauspec = .

        use data, replace

    }

else {

    use data, replace

    capture xtlogit true sens spec, nocons || study: sens spec if thresh == `i' & nm>1, nocons binomial(n)
intpoints(1) variance

    local sens = _b[sens]

    local seSens = _se[sens]

    local speci = _b[spec]

    local seSpeci = _se[spec]

    local lntausens = _b[lntausens]

    local lntauspec = _b[lntauspec]

}

local dataset = `h'

local ite = `i'

post `ests' (`dataset') (`ite') (`sens') (`seSens') (`speci') (`seSpeci') (`lntausens') (`lntauspec')

}

}

}

postclose `ests'

```

```
use `estimates_file', replace
save ID-MA, replace
**Imputed results now saved
```

APPENDIX D

APPENDIX D.1: Publication from Chapter 2

Breast Cancer Res Treat
DOI 10.1007/s10549-011-1818-2

REVIEW

A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance

Catherine Meads · Ikhtlaq Ahmed ·
Richard D. Riley

Received: 30 September 2011 / Accepted: 1 October 2011
© Springer Science+Business Media, LLC. 2011

Abstract A risk prediction model is a statistical tool for estimating the probability that a currently healthy individual with specific risk factors will develop a condition in the future such as breast cancer. Reliably accurate prediction models can inform future disease burdens, health policies and individual decisions. Breast cancer prediction models containing modifiable risk factors, such as alcohol consumption, BMI or weight, condom use, exogenous hormone use and physical activity, are of particular interest to women who might be considering how to reduce their risk of breast cancer and clinicians developing health policies to reduce population incidence rates. We performed a systematic review to identify and evaluate the performance of prediction models for breast cancer that contain modifiable factors. A protocol was developed and a sensitive search in databases including MEDLINE and EMBASE was conducted in June 2010. Extensive use was made of reference lists. Included were any articles proposing or validating a breast cancer prediction model in a general female population, with no language restrictions. Duplicate data extraction and quality assessment were conducted. Results were summarised qualitatively, and where possible meta-analysis of model performance statistics was undertaken. The systematic review found 17 breast cancer models, each containing a different but often overlapping

set of modifiable and other risk factors, combined with an estimated baseline risk that was also often different. Quality of reporting was generally poor, with characteristics of included participants and fitted model results often missing. Only four models received independent validation in external data, most notably the 'Gail 2' model with 12 validations. None of the models demonstrated consistently outstanding ability to accurately discriminate between those who did and those who did not develop breast cancer. For example, random-effects meta-analyses of the performance of the 'Gail 2' model showed the average *C* statistic was 0.63 (95% CI 0.59–0.67), and the expected/observed ratio of events varied considerably across studies (95% prediction interval for *E/O* ratio when the model was applied in practice was 0.75–1.19). There is a need for models with better predictive performance but, given the large amount of work already conducted, further improvement of existing models based on conventional risk factors is perhaps unlikely. Research to identify new risk factors with large additionally predictive ability is therefore needed, alongside clearer reporting and continual validation of new models as they develop.

Keywords Breast cancer · Systematic review · Prediction models

C. Meads (✉)
Centre for Primary Care and Public Health, Barts and The
London School of Medicine and Dentistry, Queen Mary
University of London, Yvonne Carter Building, 58 Turner St,
Whitechapel, London E1 2AB, UK
e-mail: c.meads@qmul.ac.uk

I. Ahmed · R. D. Riley
Unit of Public Health, Epidemiology and Biostatistics,
University of Birmingham, B15 2TT Birmingham, UK

Background

A risk prediction model is a statistical tool for estimating the probability that a currently healthy individual with specific risk factors (e.g. age, menopausal status) will develop a future condition, such as breast cancer, within a certain time period (such as within 5 years or lifetime). Risk models combine the baseline risk of developing the

Published online: 27 October 2011

 Springer

RESEARCH ARTICLE

Open Access

Developing and validating risk prediction models in an individual participant data meta-analysis

Ikhlaaq Ahmed¹, Thomas PA Debray², Karel GM Moons² and Richard D Riley^{3*}

Abstract

Background: Risk prediction models estimate the risk of developing future outcomes for individuals based on one or more underlying characteristics (predictors). We review how researchers develop and validate risk prediction models within an individual participant data (IPD) meta-analysis, in order to assess the feasibility and conduct of the approach.

Methods: A qualitative review of the aims, methodology, and reporting in 15 articles that developed a risk prediction model using IPD from multiple studies.

Results: The IPD approach offers many opportunities but methodological challenges exist, including: unavailability of requested IPD, missing patient data and predictors, and between-study heterogeneity in methods of measurement, outcome definitions and predictor effects. Most articles develop their model using IPD from *all* available studies and perform only an internal validation (on the same set of data). Ten of the 15 articles did not allow for any study differences in baseline risk (intercepts), potentially limiting their model's applicability and performance in some populations. Only two articles used external validation (on different data), including a novel method which develops the model on all but one of the IPD studies, tests performance in the excluded study, and repeats by rotating the omitted study.

Conclusions: An IPD meta-analysis offers unique opportunities for risk prediction research. Researchers can make more of this by allowing separate model intercept terms for each study (population) to improve generalisability, and by using 'internal-external cross-validation' to simultaneously develop and validate their model. Methodological challenges can be reduced by prospectively planned collaborations that share IPD for risk prediction.

Keywords: Meta-analysis, Prognostic factor, Prognosis, Individual participant (patient) data, Review, Reporting

Background

One of the cornerstones of health and clinical research is to identify individuals who have a high risk of developing an adverse outcome over a specific time period, so that they can be targeted for early preventative strategies and possibly treatment. For example individuals who are seemingly healthy but are found to have a high risk of developing cardiovascular disease could be recommended to modify their lifestyle and behaviour (e.g. smoking, exercise, eating habits) to reduce their future risk. They may also be prioritised for clinical investigation, which could lead to early diagnosis of an underlying condition (e.g. diabetes, high blood pressure) and preventative treatment (e.g. statins or aspirin) to manage it.

For this purpose of prognostic risk assessments there is a growing interest in *risk prediction modelling* [1-3] where a statistical model is used to estimate the risk of future outcomes for individuals based on one or more underlying characteristics. When considering future outcomes in patients, a risk prediction model is often referred to as a *prognostic model* (typically used for outcome risk for a defined disease) or more generally a *clinical prediction model* (used for both diseased or non-diseased settings). Similarly the word 'model' is often replaced with 'score', 'tool', 'index', or 'rule'. However, the same principle remains: to accurately predict the risk of future occurrence of an outcome in an individual by utilising the values or levels of multiple individual characteristics. We refer here to such characteristics simply as predictors, but they are also termed prognostic factors, risk factors, prognostic variables, and prognostic markers [4]. They often include

* Correspondence: r.d.riley@bham.ac.uk

³School of Health and Population Sciences, Public Health Building, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK

Full list of author information is available at the end of the article



© 2014 Ahmed et al.; licensee BioMed Central Ltd. This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Summarising and validating test accuracy results across multiple studies for use in clinical practice

Richard D. Riley,^{a,*†} Ikhlmaq Ahmed,^b
Thomas P. A. Debray,^c Brian H. Willis,^d J. Pieter Noordzij,^e
Julian P.T. Higgins^f and Jonathan J. Deeks^d

Following a meta-analysis of test accuracy studies, the translation of summary results into clinical practice is potentially problematic. The sensitivity, specificity and positive (PPV) and negative (NPV) predictive values of a test may differ substantially from the average meta-analysis findings, because of heterogeneity. Clinicians thus need more guidance: given the meta-analysis, is a test likely to be useful in new populations, and if so, how should test results inform the probability of existing disease (for a diagnostic test) or future adverse outcome (for a prognostic test)? We propose ways to address this. Firstly, following a meta-analysis, we suggest deriving prediction intervals and probability statements about the potential accuracy of a test in a new population. Secondly, we suggest strategies on how clinicians should derive post-test probabilities (PPV and NPV) in a new population based on existing meta-analysis results and propose a cross-validation approach for examining and comparing their calibration performance. Application is made to two clinical examples. In the first example, the joint probability that both sensitivity and specificity will be >80% in a new population is just 0.19, because of a low sensitivity. However, the summary PPV of 0.97 is high and calibrates well in new populations, with a probability of 0.78 that the true PPV will be at least 0.95. In the second example, post-test probabilities calibrate better when tailored to the prevalence in the new population, with cross-validation revealing a probability of 0.97 that the observed NPV will be within 10% of the predicted NPV. © 2015 The Authors. *Statistics in Medicine* Published by John Wiley & Sons Ltd.

Keywords: meta-analysis; test accuracy; prognostic; diagnostic; calibration; discrimination

1. Introduction

Test accuracy studies aim to evaluate the performance of a candidate medical test for either diagnosing the presence of a particular clinical condition ('diagnostic test') or identifying those likely to experience a particular outcome in the future ('prognostic test'). Such tests include measurable variables such as biomarkers, blood pressure and temperature, or may reflect a clinician's judgement after a physical examination or imaging result. When multiple studies evaluate the performance of a potential diagnostic or prognostic test, meta-analysis methods can synthesise the study results to help establish if and how

^aResearch Institute of Primary Care and Health Sciences, Keele University, Staffordshire ST5 5BG, U.K.

^bMRC Hub for Trials Methodology Research, School of Health and Population Sciences, University of Birmingham, Birmingham, U.K.

^cJulius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands

^dSchool of Health and Population Sciences, University of Birmingham, Birmingham, U.K.

^eDepartment of Otolaryngology – Head & Neck Surgery, Boston Medical Center, Boston University – School of Medicine, Boston, MA, U.S.A.

^fSchool of Social and Community Medicine, University of Bristol, Bristol, U.K.

*Correspondence to: Richard Riley, Research Institute of Primary Care and Health Sciences, Keele University, Staffordshire ST5 5BG, U.K.

†E-mail: r.riley@keele.ac.uk

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

APPENDIX D.4: Paper submitted for publication from work in Chapter 6

Riley et al. *Systematic Reviews* 2015, 4:12
<http://www.systematicreviewsjournal.com/content/4/1/12>



METHODOLOGY

Open Access

Meta-analysis of test accuracy studies: an exploratory method for investigating the impact of missing thresholds

Richard D Riley^{1*}, Ikhlaaq Ahmed³, Joie Ensor², Yemisi Takwoingi², Amanda Kirkham², R Katie Morris^{4,5}, J Pieter Noordzij⁶ and Jonathan J Deeks²

Abstract

Background: Primary studies examining the accuracy of a continuous test evaluate its sensitivity and specificity at one or more thresholds. Meta-analysts then usually perform a separate meta-analysis for each threshold. However, the number of studies available for each threshold is often very different, as primary studies are inconsistent in the thresholds reported. Furthermore, of concern is selective reporting bias, because primary studies may be less likely to report a threshold when it gives low sensitivity and/or specificity estimates. This may lead to biased meta-analysis results. We developed an exploratory method to examine the potential impact of missing thresholds on conclusions from a test accuracy meta-analysis.

Methods: Our method identifies studies that contain missing thresholds bounded between a pair of higher and lower thresholds for which results are available. The bounded missing threshold results (two-by-two tables) are then imputed, by assuming a linear relationship between threshold value and each of logit-sensitivity and logit-specificity. The imputed results are then added to the meta-analysis, to ascertain if original conclusions are robust. The method is evaluated through simulation, and application made to 13 studies evaluating protein:creatinine ratio (PCR) for detecting proteinuria in pregnancy with 23 different thresholds, ranging from one to seven per study.

Results: The simulation shows the imputation method leads to meta-analysis estimates with smaller mean-square error. In the PCR application, it provides 50 additional results for meta-analysis and their inclusion produces lower test accuracy results than originally identified. For example, at a PCR threshold of 0.16, the summary specificity is 0.80 when using the original data, but 0.66 when also including the imputed data. At a PCR threshold of 0.25, the summary sensitivity is reduced from 0.95 to 0.85 when additionally including the imputed data.

Conclusions: The imputation method is a practical tool for researchers (often non-statisticians) to explore the potential impact of missing threshold results on their meta-analysis conclusions. Software is available to implement the method. In the PCR example, it revealed threshold results are vulnerable to the missing data, and so stimulates the need for advanced statistical models or, preferably, individual patient data from primary studies.

Keywords: Meta-analysis, Diagnostic test, Multiple thresholds, Imputation, Missing data, Sensitivity analysis

* Correspondence: r.riley@kcl.ac.uk

¹Research Institute of Primary Care and Health Sciences, Keele University, Staffordshire ST5 5BG, UK

Full list of author information is available at the end of the article



© 2015 Riley et al.; licensee BioMed Central. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.

List of References

1. Moons, K.G., et al., *Prognosis and prognostic research: what, why, and how?* Bmj, 2009. **338**: p. b375.
2. Moons, K.G., et al., *Prognosis and prognostic research: application and impact of prognostic models in clinical practice.* Bmj, 2009. **338**: p. b606.
3. Altman, D.G., et al., *Prognosis and prognostic research: validating a prognostic model.* BMJ, 2009. **338**: p. b605.
4. Royston, P., et al., *Prognosis and prognostic research: developing a prognostic model.* British Medical Journal, 2009. **338**: p. b604 (1373-1377).
5. Steyerberg, E.W., et al., *Predicting outcome after traumatic brain injury: Development and international validation of prognostic scores based on admission characteristics.* PLoS Medicine, 2008. **5**(8): p. 1251-1261.
6. Wyatt, J. and D.G. Altman, *Commentary: Prognostic models: clinically useful or quickly forgotten?* BMJ, 1995. **311**: p. 1539-1541.
7. Reilly, B.M. and A.T. Evans, *Translating clinical research into clinical practice: impact of using prediction rules to make decisions.* Ann Intern Med, 2006. **144**(3): p. 201-9.
8. Hemingway, H., et al., *Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes.* Bmj, 2013. **346**.
9. Moons KG, et al., *Prognosis and prognostic research: what, why, and how?* BMJ, 2009. **338**(b375).
10. Riley, R.D., et al., *Prognosis Research Strategy (PROGRESS) 2: Prognostic Factor Research.* PLoS Medicine, 2013. **10**(2): p. e1001380.
11. Moons, K.G.M., et al., *Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker.* Heart, 2012. **98**(9): p. 683-690.
12. Royston P, et al., *Prognosis and prognostic research: developing a prognostic model.* BMJ, 2009. **338**(b604).
13. Steyerberg, E.W., *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating.* 2008: Springer.
14. Altman DG, et al., *Prognosis and prognostic research: validating a prognostic model.* BMJ, 2009. **338**(b605).
15. Altman, D.G. and P. Royston, *The cost of dichotomising continuous variables.* Vol. 332. 2006. 1080.
16. Ahmed, I., et al., *Developing and validating risk prediction models in an individual participant data meta-analysis.* BMC Medical Research Methodology, 2014. **14**(1): p. 3.
17. COX, D.R., *Partial likelihood.* Biometrika, 1975. **62**(2): p. 269-276.
18. Solal-Céligny, P., et al., *Follicular Lymphoma International Prognostic Index.* Vol. 104. 2004. 1258-1265.
19. Steyerberg, E.W., et al., *Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research.* PLoS Med, 2013. **10**(2): p. e1001381.
20. Steyerberg, E.W., et al., *Assessing the performance of prediction models: a framework for traditional and novel measures.* Epidemiology. **2010 Jan**; **21**(1): p. 128-38.
21. Uno, H., et al., *On the C-statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data.* Statistics in medicine, 2011. **30**(10): p. 1105-1117.
22. Cochrane, *Diagnostic Test Accuracy Working Group.* 2014.
23. Hemingway Pippa and B. Nic, *What is a systematic review? What is...? series,* 2009.

24. Rothstein HR, Sutton AJ, and B.M. (eds). *Publication Bias in Meta-Analysis*. Wiley, 2006.
25. Dubben HH and B.-B. HP, *Systematic review of publication bias in studies on publication bias*. *BMJ*, 2005. **331**: p. 433-434.
26. GV, G., *Primary, secondary and meta-analysis of research*. *Educational Researcher*, 1976. **5**: p. 3-8.
27. van Klaveren, D., et al., *Assessing discriminative ability of risk models in clustered data*. *BMC Medical Research Methodology*, 2014. **14**(1): p. 5.
28. Crombie Iain K and D.H. TO, *What is meta-analysis? What is...? series*, 2009.
29. Moher D, L.A., Tetzlaff J, Altman DG, *Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement*. *BMJ*, 2009. **339**(b2535).
30. Moher D, Cook DJ, and E.S.e. al., *Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement*. *Quality of Reporting of Meta-analyses*. *Lancet*, 1999. **354**: p. 1896-1900.
31. Sutton AJ, *Meta-Analysis 1: introduction to principles and practice, lecture*. 2010.
32. Riley, R.D., J.P.T. Higgins, and J.J. Deeks, *Interpretation of random effects meta-analyses*. *Bmj*, 2011. **342**.
33. Riley RD, Lambert PC, and A.-Z. G, *Meta-analysis of individual participant data: rationale, conduct, and reporting*. *BMJ*, 2010. **340**(feb05_1): p. c221 - c221.
34. Simmonds MC, Clarke MJ, and S. LA, *Meta-analysis of individual patient data from randomized trials: a review of methods used in practice*. *Clinical Trials*, 2005. **2**: p. 209-217.
35. Debray, T.P., et al., *Individual participant data meta-analysis for a binary outcome: one-stage or two-stage?* *PloS one*, 2013. **8**(4): p. e60650.
36. Stewart LA and P. MK, *Meta-analysis of the literature or of individual patient data: is there a difference?* . *Lancet*, 1993. **341**(8842): p. 418-422.
37. Simmonds MC, et al., *Meta-analysis of individual patient data from randomized trials: a review of methods used in practice*. *Clinical Trials*, 2005. **2**: p. 209-217.
38. Riley RD, et al., *Meta-analysis of diagnostic test studies using individual patient data and aggregate data*. *Stat Med*, 2008. **27**(29): p. 6111-36.
39. Higgins JP, et al., *Meta-analysis of continuous outcome data from individual patients*. . *Stat Med*, 2001. **20**(15): p. 2219-41.
40. Stewart, L.A. and J.F. Tierney, *To IPD or not to IPD?: Advantages and Disadvantages of Systematic Reviews Using Individual Patient Data*. *Evaluation & the Health Professions*, 2002. **25**(1): p. 76-97.
41. Meads, C., I. Ahmed, and R. Riley, *A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance*. *Breast Cancer Research and Treatment*: p. 1-13.
42. Riley, R.D., et al., *Summarising and validating test accuracy results across multiple studies for use in clinical practice*. *Statistics in Medicine*, 2015. **34**(13): p. 2081-2103.
43. Riley, R., et al., *Meta-analysis of test accuracy studies: an exploratory method for investigating the impact of missing thresholds*. *Systematic Reviews*, 2015. **4**(1): p. 12.
44. breastcancer.org,
http://www.breastcancer.org/symptoms/understand_bc/what_is_bc.jsp. accessed 08/02/2012.
45. cancerresearchuk.org,
<http://info.cancerresearchuk.org/cancerstats/types/breast/incidence/>. accessed on 08/02/2012.
46. Altman, D.G., *Prognostic Models: A Methodological Framework and Review of Models for Breast Cancer*. *Cancer Investigation*, 2009. **27**(3): p. 235-243.

47. Tice, J.A., et al., *Using Clinical Factors and Mammographic Breast Density to Estimate Breast Cancer Risk: Development and Validation of a New Predictive Model*. Annals of Internal Medicine, 2008. **148**(5): p. 337-347.
48. DerSimonian R, L.N., *Meta-analysis in clinical trials*. Control Clin Trials, 1986(7): p. 177-188.
49. Corporation., S., *Statistical software release 12.1*. College Station, Texas, 2011.
50. GN, A., *Breast cancer risk assessments to barrier contraception exposure. A new approach*. 2009. **30**(1): p. 217-232.
51. Colditz, G.A. and B. Rosner, *Cumulative Risk of Breast Cancer to Age 70 Years According to Risk Factor Status: Data from the Nurses' Health Study*. American Journal of Epidemiology, 2000. **152**(10): p. 950-964.
52. Cook, N.R., et al., *Mammographic Screening and Risk Factors for Breast Cancer*. American Journal of Epidemiology, 2009. **170**(11): p. 1422-1432.
53. Gail, M.H., et al., *Projecting Individualized Probabilities of Developing Breast Cancer for White Females Who Are Being Examined Annually*. Journal of the National Cancer Institute, 1989. **81**(24): p. 1879-1886.
54. Tyrer, J., S.W. Duffy, and J. Cuzick, *A breast cancer prediction model incorporating familial and personal risk factors*. Statistics in Medicine, 2004. **23**(7): p. 1111-1130.
55. Wacholder, S., et al., *Performance of Common Genetic Variants in Breast-Cancer Risk Models*. New England Journal of Medicine, 2010. **362**(11): p. 986-993.
56. Amir, E., et al., *Evaluation of breast cancer risk assessment packages in the family history evaluation and screening programme*. Journal of Medical Genetics, 2003. **40**(11): p. 807-814.
57. Bondy, M.L., et al., *Validation of a Breast Cancer Risk Assessment Model in Women With a Positive Family History*. Journal of the National Cancer Institute, 1994. **86**(8): p. 620-625.
58. Costantino, J.P., et al., *Validation Studies for Models Projecting the Risk of Invasive and Total Breast Cancer Incidence*. Journal of the National Cancer Institute, 1999. **91**(18): p. 1541-1548.
59. Rockhill, B., et al., *Validation of the Gail et al. Model of Breast Cancer Risk Prediction and Implications for Chemoprevention*. Journal of the National Cancer Institute, 2001. **93**(5): p. 358-366.
60. Spiegelman, D., et al., *Validation of the Gail et al. Model for Predicting Individual Breast Cancer Risk*. Journal of the National Cancer Institute, 1994. **86**(8): p. 600-607.
61. Schonfeld, S.J., et al., *Effect of Changing Breast Cancer Incidence Rates on the Calibration of the Gail Model*. Journal of Clinical Oncology, 2010. **28**(14): p. 2411-2417.
62. Ulusoy, C., et al., *Applicability of the Gail model for breast cancer risk assessment in Turkish female population and evaluation of breastfeeding as a risk factor*. Breast Cancer Research and Treatment, 2010. **120**(2): p. 419-424.
63. Rockhill, B., et al., *Breast cancer risk prediction with a log-incidence model: evaluation of accuracy*. Journal of clinical epidemiology, 2003. **56**(9): p. 856-861.
64. Viallon, V., et al., *How to evaluate the calibration of a disease risk prediction tool*. Statistics in Medicine, 2009. **28**(6): p. 901-916.
65. Barlow, W.E., et al., *Prospective Breast Cancer Risk Prediction Model for Women Undergoing Screening Mammography*. Journal of the National Cancer Institute, 2006. **98**(17): p. 1204-1214.
66. Boyle, P., et al., *Contribution of three components to individual cancer risk predicting breast cancer risk in Italy*. European Journal of Cancer Prevention, 2004. **13**(3): p. 183-191.

67. Chen, J., et al., *Projecting Absolute Invasive Breast Cancer Risk in White Women With a Model That Includes Mammographic Density*. Journal of the National Cancer Institute, 2006. **98**(17): p. 1215-1226.
68. Decarli, A., et al., *Gail Model for Prediction of Absolute Risk of Invasive Breast Cancer: Independent Evaluation in the Florence–European Prospective Investigation Into Cancer and Nutrition Cohort*. Journal of the National Cancer Institute, 2006. **98**(23): p. 1686-1693.
69. Gail, M.H., et al., *Projecting Individualized Absolute Invasive Breast Cancer Risk in African American Women*. Journal of the National Cancer Institute, 2007. **99**(23): p. 1782-1792.
70. Novotny, J., et al., *Breast cancer risk assessment in the Czech female population – an adjustment of the original Gail model*. Breast Cancer Research and Treatment, 2006. **95**(1): p. 29-35.
71. Rosner, B., G.A. Colditz, and W.C. Willett, *Reproductive Risk Factors in a Prospective Study of Breast Cancer: The Nurses' Health Study*. American Journal of Epidemiology, 1994. **139**(8): p. 819-835.
72. Rosner, B. and G.A. Colditz, *Nurses' Health Study: Log-Incidence Mathematical Model of Breast Cancer Incidence*. Journal of the National Cancer Institute, 1996. **88**(6): p. 359-364.
73. Rosner, B., et al., *Risk prediction models with incomplete data with application to prediction of estrogen receptor-positive breast cancer: prospective data from the Nurses' Health Study*. Breast Cancer Research, 2008. **10**(4): p. R55.
74. Tice, J., et al., *Mammographic Breast Density and the Gail Model for Breast Cancer Risk Prediction in a Screening Population*. Breast Cancer Research and Treatment, 2005. **94**(2): p. 115-122.
75. Institute, N.C., *Seer database*. <http://seer.cancer.gov/data/index.html>.
76. *First results from the International Breast Cancer Intervention Study (IBIS-I): a randomised prevention trial*. The Lancet, 2002. **360**(9336): p. 817-824.
77. Pike, M.C., et al., *'Hormonal' risk factors, 'breast tissue age' and the age-incidence of breast cancer*. Nature, 1983. **303**(5920): p. 767-770.
78. Rucker, G., et al., *Undue reliance on I2 in assessing heterogeneity may mislead*. BMC Medical Research Methodology, 2008. **8**(1): p. 79.
79. Higgins JPT, T.S., Spiegelhalter DJ., *A re-evaluation of random-effects meta-analysis*. J R Stat Soc Ser A, 2009(172): p. 137-59.
80. Mallett, S., et al., *Reporting performance of prognostic models in cancer: a review*. BMC Medicine, 2010. **8**(1): p. 21.
81. Collins, G.S., et al., *Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement*. Vol. 350. 2015.
82. Debray, T.P.A., et al., *A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis*. Statistics in Medicine, 2013. **32**(18): p. 3158-3180.
83. Riley, R.D., M.C. Simmonds, and M.P. Look, *Evidence synthesis combining individual patient data and aggregate data: a systematic review identified current practice and possible methods*. Journal of clinical epidemiology, 2007. **60**(5): p. 431.e1-431.e12.
84. Abo-Zaid, G., W. Sauerbrei, and R. Riley, *Individual participant data meta-analysis of prognostic factor studies: state of the art? BMC Medical Research Methodology*, 2012. **12**(1): p. 56.
85. Pagliaro, L., et al., *Interferon- α for chronic hepatitis C: An analysis of pretreatment clinical predictors of response*. Hepatology, 1994. **19**(4): p. 820-828.
86. Heffner, J.E., P.J. Nietert, and C. Barbieri, *Pleural Fluid pH as a Predictor of Pleurodesis Failure*Analysis of Primary Data*. CHEST Journal, 2000. **117**(1): p. 87-95.

87. Raboud J. M., S.R., J. S. G. Montaner., *Predicting HIV RNA Virologic Outcome at 52-Weeks Follow-Up in Antiretroviral Clinical Trials*. JAIDS Journal of Acquired Immune Deficiency Syndromes, 2000. **24**: p. 433-439.
88. Terwee, C.B., et al., *Pooling of prognostic studies in cancer of the pancreatic head and periampullary region: the Triple-P study*. European Journal of Surgery, 2000. **166**(9): p. 706-712.
89. Chau, I., et al., *Multivariate Prognostic Factor Analysis in Locally Advanced and Metastatic Esophago-Gastric Cancer—Pooled Analysis From Three Multicenter, Randomized, Controlled Trials Using Individual Patient Data*. Journal of Clinical Oncology, 2004. **22**(12): p. 2395-2403.
90. Horn, J., et al., *Identification of Patients at Risk for Ischaemic Cerebral Complications After Carotid Endarterectomy with TCD Monitoring*. European Journal of Vascular and Endovascular Surgery, 2005. **30**(3): p. 270-274.
91. Nieder, C., et al., *Proposal of human spinal cord reirradiation dose based on collection of data from 40 patients*. International Journal of Radiation Oncology*Biophysics, 2005. **61**(3): p. 851-855.
92. Patel., A., *An evaluation of metabolic risks for coronary death in the Asia Pacific region*. Diabetes research and clinical practice, 2006. **74**(3): p. 274-281.
93. Sylvester, R.J., et al., *Predicting Recurrence and Progression in Individual Patients with Stage Ta T1 Bladder Cancer Using EORTC Risk Tables: A Combined Analysis of 2596 Patients from Seven EORTC Trials*. European Urology, 2006. **49**(3): p. 466-477.
94. Noordzij, J.P., et al., *Early Prediction of Hypocalcemia after Thyroidectomy using Parathyroid Hormone: An Analysis of Pooled Individual Patient Data from Nine Observational Studies*. Journal of the American College of Surgeons.205(6)(pp 748-754), 2007.Date of Publication: Dec 2007., 2007(6): p. 748-754.
95. Rovers, M.M., et al., *Predictors of Pain and/or Fever at 3 to 7 Days for Children With Acute Otitis Media Not Treated Initially With Antibiotics: A Meta-analysis of Individual Patient Data*. Pediatrics, 2007. **119**(3): p. 579-585.
96. Schaich, M., et al., *Prognosis of acute myeloid leukemia patients up to 60 years of age exhibiting trisomy 8 within a non-complex karyotype: individual patient data-based meta-analysis of the German Acute Myeloid Leukemia Intergroup*. Haematologica, 2007. **92**(6): p. 763-770.
97. Fowkes, F.G.R.M., G.D. ; Butcher, I., *Ankle brachial index combined with Framingham risk score to predict cardiovascular events and mortality - A meta-analysis*. JAMA, 2008. **300**(2): p. 197-208.
98. Steyerberg EW, et al., *Predicting outcome after traumatic brain injury: Development and international validation of prognostic scores based on admission characteristics*. PLoS Medicine, 2008. **5**(8): p. 1251-1261.
99. Yap, Y.G., et al., *Potential demographic and baselines variables for risk stratification of high-risk post-myocardial infarction patients in the era of implantable cardioverter-defibrillator — A prognostic indicator*. International Journal of Cardiology, 2008. **126**(1): p. 101-107.
100. Ahmed, I., A.J. Sutton, and R.D. Riley, *Assessment of publication bias, selection bias, and unavailable data in meta-analyses using individual participant data: a database survey*. Bmj, 2012. **344**.
101. Royston P, Parmar MKB, and S. R., *Construction and validation of a prognostic model across several studies, with an application in superficial bladder cancer*. Statistics in Medicine, 2004. **23**: p. 907-926.
102. Abo-Zaid, G., et al., *Individual participant data meta-analyses should not ignore clustering*. J Clin Epidemiol, 2013. **66**(8): p. 865-873 e4.

103. Burton, A., et al., *The design of simulation studies in medical statistics*. Statistics in Medicine, 2006. **25**(24): p. 4279-4292.
104. May, M., et al., *Development and validation of a prognostic model for survival time data: application to prognosis of HIV positive patients treated with antiretroviral therapy*. Statistics in Medicine, 2004. **23**(15): p. 2375-2398.
105. Teschendorff, A.E., et al., *A consensus prognostic gene expression classifier for ER positive breast cancer*. Genome biology, 2006. **7**(10): p. R101.
106. Abrahantes, J.C., et al., *Comparison of different estimation procedures for proportional hazards model with random effects*. Computational statistics & data analysis, 2007. **51**(8): p. 3913-3930.
107. May, M., et al., *A coronary heart disease risk model for predicting the effect of potent antiretroviral therapy in HIV-1 infected men*. International journal of epidemiology, 2007. **36**(6): p. 1309-1318.
108. Kaptoge, S., et al., *Prediction of incident hip fracture risk by femur geometry variables measured by hip structural analysis in the study of osteoporotic fractures*. Journal of Bone and Mineral Research, 2008. **23**(12): p. 1892-1904.
109. Fibrinogen, S.C., *Measures to assess the prognostic ability of the stratified Cox proportional hazards model*. Statistics in Medicine, 2009. **28**(3): p. 389.
110. Kalogeropoulos, A.P., et al., *Utility of the Seattle Heart Failure Model in patients with advanced heart failure*. Journal of the American College of Cardiology, 2009. **53**(4): p. 334-342.
111. Legrand, C., et al., *Validation of prognostic indices using the frailty model*. Lifetime data analysis, 2009. **15**(1): p. 59-78.
112. Wood, A.M. and P. Greenland, *Evaluating the prognostic value of new cardiovascular biomarkers*. Disease Markers, 2009. **26**(5): p. 199-207.
113. Di Maio, M., et al., *Clinical assessment of patients with advanced non-small-cell lung cancer eligible for second-line chemotherapy: a prognostic score from individual data of nine randomised trials*. European Journal of Cancer, 2010. **46**(4): p. 735-743.
114. Friis-Møller, N., et al., *Predicting the risk of cardiovascular disease in HIV-infected patients: the data collection on adverse effects of anti-HIV drugs study*. European Journal of Cardiovascular Prevention & Rehabilitation, 2010. **17**(5): p. 491-501.
115. May, M., et al., *Prognosis of HIV-1 infected patients starting antiretroviral therapy in sub-Saharan Africa: a collaborative analysis of scale-up programmes*. Lancet, 2010. **376**(9739): p. 449.
116. Riley, R.D., *Commentary: Like it and lump it? Meta-analysis using individual participant data*. International journal of epidemiology, 2010. **39**(5): p. 1359-1361.
117. Riley RD, L.P., Abo-Zaid G., *Meta-analysis of individual participant data: conduct, rationale and reporting*. BMJ, 2010. **340**: p. c221.
118. Royston, P., M.K.B. Parmar, and D.G. Altman, *External validation and updating of a prognostic survival model*. 2010, Oxford Research Report.
119. Buuren, S. and K. Groothuis-Oudshoorn, *MICE: Multivariate imputation by chained equations in R*. Journal of Statistical Software, 2011. **45**(3).
120. Cai, T., et al., *Robust Prediction of t-Year Survival with Data from Multiple Studies*. Biometrics, 2011. **67**(2): p. 436-444.
121. Sauerbrei, W. and P. Royston, *A new strategy for meta-analysis of continuous covariates in observational studies*. Statistics in Medicine, 2011. **30**(28): p. 3341-3360.
122. Ambler, G., S. Seaman, and R. Omar, *An evaluation of penalised survival methods for developing prognostic models with rare events*. Statistics in Medicine, 2012.
123. Barker, A.L., et al., *Mobility has a non-linear association with falls risk among people in residential aged care: an observational study*. Journal of Physiotherapy, 2012. **58**(2): p. 117.

124. Petoumenos, K., et al., *Predicting the short-term risk of diabetes in HIV-positive patients: the Data Collection on Adverse Events of Anti-HIV Drugs (D: A: D) study*. Journal of the International AIDS Society, 2012. **15**(2).
125. Phillips, R.S., et al., *Predicting infectious complications in neutropenic children and young people with cancer (IPD protocol)*. Systematic Reviews, 2012. **1**: p. 8.
126. Van Buuren, S., *Flexible imputation of missing data*. 2012: Chapman & Hall.
127. Farooq, V., et al., *Prediction of 1-year mortality in patients with acute coronary syndromes undergoing percutaneous coronary intervention: validation of the logistic clinical syntax (synergy between percutaneous coronary interventions with taxus and cardiac surgery) score*. JACC: Cardiovascular Interventions, 2013. **6**(7): p. 737-745.
128. Farooq, V., et al., *Widening clinical applications of the SYNTAX Score*. Heart, 2013.
129. Raichand, S., et al., *Protocol for a systematic review of the diagnostic and prognostic utility of tests currently available for the detection of aspirin resistance in patients with established cardiovascular or cerebrovascular disease*. Systematic reviews, 2013. **2**(1): p. 16.
130. Pennells, L., et al., *Assessing risk prediction models using individual participant data from multiple studies*. American journal of epidemiology, 2014. **179**(5): p. 621-632.
131. Davies, M.-A., et al., *Prognosis of Children With HIV-1 Infection Starting Antiretroviral Therapy in Southern Africa: A Collaborative Analysis of Treatment Programs*. The Pediatric infectious disease journal, 2014. **33**(6): p. 608-616.
132. Phillips, R.S., *Optimizing risk predictive strategies in febrile neutropenic episodes in children and young people undergoing treatment for malignant disease*. 2014, University of York.
133. Sene, M., *Développement d'outils pronostiques dynamiques dans le cancer de la prostate localisé traité par radiothérapie*. 2013, Bordeaux 2.
134. Rodríguez Mendiola, N.M., *Predicción de mortalidad en pacientes en hemodiálisis: diseño y validación de un índice pronóstico*. 2013.
135. Putter, H., M. Fiocco, and T. Stijnen, *Meta-Analysis of Diagnostic Test Accuracy Studies with Multiple Thresholds using Survival Methods*. Biometrical Journal, 2010. **52**(1): p. 95-110.
136. Wikipedia, <https://en.wikipedia.org/wiki/Thyroidectomy>. accessed on 30/05/13.
137. Wikipedia, <http://en.wikipedia.org/wiki/Hypocalcaemia> accessed 24/04/12.
138. BBC, http://www.bbc.co.uk/health/physical_health/conditions/hypocalcaemia1.shtml accessed 24/04/12.
139. Riley, R., *Multivariate meta-analysis: the effect of ignoring within-study correlation*. JRSS Series A, 2009. **172**(4): p. 789-811.
140. Kirkham, J.J., R.D. Riley, and P.R. Williamson, *A multivariate meta-analysis approach for reducing the impact of outcome reporting bias in systematic reviews*. Statistics in Medicine, 2012. **31**(20): p. 2179-2195.
141. Lam, A. and P.D. Kerr, *Parathyroid Hormone: An Early Predictor of Postthyroidectomy Hypocalcemia*. The Laryngoscope, 2003. **113**(12): p. 2196-2200.
142. Lo CY, L.J., Tam SC, *Applicability of intraoperative parathyroid hormone assay during thyroidectomy*. Ann Surg, 2002. **236**: p. 564-569.
143. Lombardi CP, R.M., Princi P, et al., *Early prediction of postthyroidectomy hypocalcemia by one single iPTH measurement*. Surgery, 2004. **136**: p. 1236-1240.
144. McLeod IK, A.C., Noordzij JP, et al., *The use of rapid parathyroid hormone assay in predicting postoperative hypocalcemia after total or completion thyroidectomy*. Thyroid, 2006. **16**: p. 259-265.
145. Warren, F.M., et al., *Intraoperative Parathyroid Hormone Levels in Thyroid and Parathyroid Surgery*. The Laryngoscope, 2002. **112**(10): p. 1866-1870.

146. Warren, F.M., et al., *Perioperative Parathyroid Hormone Levels in Thyroid Surgery: Preliminary Report*. The Laryngoscope, 2004. **114**(4): p. 689-693.
147. Youden, W.J., *Index for rating diagnostic tests*. Cancer, 1950. **3**(1): p. 32-35.
148. Hamza, T., et al., *Multivariate random effects meta-analysis of diagnostic tests with multiple thresholds*. BMC Medical Research Methodology, 2009. **9**(1): p. 73.
149. StataCorp, *Stata Statistical Software: Release 12*. College Station, TX: StataCorp LP., 2011.
150. Chu, H. and S.R. Cole, *Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach*. Journal of clinical epidemiology, 2006. **59**(12): p. 1331-1332.
151. Hamza, T.H., H.C. van Houwelingen, and T. Stijnen, *The binomial distribution of meta-analysis was preferred to model within-study variability*. Journal of clinical epidemiology, 2008. **61**(1): p. 41-51.
152. Riley, R., et al., *An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes*. Statistics in Medicine, 2007. **26**(1): p. 78-97.
153. Haitao, C. and S.R. Cole, *Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach*. Journal of clinical epidemiology, 2006. **59**(12): p. 1331-1332.
154. Reitsma, J.B., et al., *Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews*. Journal of Clinical Epidemiology, 2005. **58**(10): p. 982-990.
155. Selvin, S., *Survival analysis for epidemiologic and medical research*. 2008: Cambridge University Press New York, NY.
156. Willis, B.H. and C.J. Hyde, *Estimating a test's accuracy using tailored meta-analysis—How setting-specific data may aid study selection*. Journal of Clinical Epidemiology, 2014. **67**(5): p. 538-546.
157. Hunink, G.M., et al., *Decision Making in Health and Medicine: Integrating Evidence and Values*. 2001: Cambridge University Press.
158. Richard D Riley, et al., *Meta-analysis of test accuracy studies with multiple and missing thresholds: a sensitivity analysis using imputation*. Submitted, 2014.
159. Simel, D.L. and P.M. Bossuyt, *Differences between univariate and bivariate models for summarizing diagnostic accuracy may not be large*. Journal of clinical epidemiology, 2009. **62**(12): p. 1292-1300.
160. Riley, R., et al., *Bivariate random-effects meta-analysis and the estimation of between-study correlation*. BMC Medical Research Methodology, 2007. **7**(1): p. 1-15.
161. Pinheiro JC, B.D., *Approximations to the Log-likelihood Function in the Nonlinear Mixed-effects Model*. Approximations to the Log-likelihood Function in the Nonlinear Mixed-effects Model., 1995(4): p. 12-35.
162. Inc, S.I., *PROC NLMIXED*. SAS Institute Inc, 1999. **Cary, NC**.
163. Jackson, D., R. Riley, and I.R. White, *Multivariate meta-analysis: Potential and promise*. Statistics in Medicine, 2011. **30**(20): p. 2481-2498.
164. Leeflang, M.M., et al., *Bias in sensitivity and specificity caused by data-driven selection of optimal cutoff values: mechanisms, magnitude, and solutions*. Clinical chemistry, 2008. **54**(4): p. 729-737.
165. Riley RD, A.I., Debray TPA, Willis BH, Noordzij JP, Higgins JPT, Deeks JJ, *Summarising and validating test accuracy results across multiple studies for use in clinical practice*. (submitted), 2015.
166. Leeflang, M.M.G., et al., *Bivariate meta-analysis of predictive values of diagnostic tests can be an alternative to bivariate meta-analysis of sensitivity and specificity*. Journal of Clinical Epidemiology. **65**(10): p. 1088-1097.
167. Group, C.P.M.

168. Burgess, S., I.R. White, and M. Resche-Rigon, *Combining multiple imputation and meta-analysis*. 2012.
169. Riley, R., et al., *Meta-analysis of test accuracy studies with multiple and missing thresholds: a multivariate-normal model*. J Biomed Biostat, 2014. **5**: p. 100196.
170. Dukic, V. and C. Gatsonis, *Meta-analysis of Diagnostic Test Accuracy Assessment Studies with Varying Number of Thresholds*. Biometrics, 2003. **59**(4): p. 936-946.
171. Putter, H., M. Fiocco, and T. Stijnen, *Meta-Analysis of Diagnostic Test Accuracy Studies with Multiple Thresholds using Survival Methods*. Biometrical Journal, 2010. **52**(1): p. 95-110.